УДК 004.056, 004.852, 004.042
DOI: 10.15593/2224-9397/2020.4.02

С.В. Корелов[1], А.М. Петров[1], Л.Ю. Ротков[2], А.А. Горбунов[2]

[1]Национальный координационный центр по компьютерным инцидентам, Москва, Россия
[2]Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского, Нижний Новгород, Россия

ОПРЕДЕЛЕНИЕ ДЛИНЫ ВЫБОРКИ В МОДЕЛИ ЭЛЕКТРОННЫХ ПИСЕМ

В условиях интенсивного развития различных сфер деятельности государства и общества использование передовых информационных технологий становится одним из наиболее важных, а часто и решающим фактором, определяющим эффективность всех уровней управления. Информационные технологии занимают прочные позиции в повседневной жизни и в различных сферах деятельности государства. Для обработки и анализа информации, прогнозирования и поддержки принятия управленческих решений различного уровня внедряют информационные системы различных видов и назначений. Функционирование практически любой организации в той или иной степени зависит от того, насколько эффективно и устойчиво функционирует ее информационная система, насколько надежно защищены ее информационные ресурсы от различных угроз безопасности информации, одной из которых является спам. Стоит отметить, что было совершено множество попыток раз и навсегда решить проблему его обнаружения. В данной предметной области постоянно ведутся исследования. По их результатам предлагаются и реализуются на практике различные подходы, которые по большей части основываются на хорошо зарекомендовавшей себя базовой статистической обработке различных параметров электронных писем. При этом не вполне учитывается содержание электронных писем, что приводит к росту числа ошибок первого рода. Для решения данного вопроса ранее авторами предложена модель электронных писем, учитывающая содержание электронных писем, которое зачастую меняется в зависимости от выполняемых пользователями задач и их информационных потребностей. Ее ключевой особенностью является то, что она оперирует с преобразованными в числовой вектор данными, полученными из исходных текстов электронных писем. В качестве параметров модели электронных писем, оказывающих влияние на выделение текстовых отрезков писем, являющихся отражением их отличительных признаков, авторами выделены: количество числовых кодов, сопоставляемых символам текста, в функции преобразования писем в числовой вектор; шаг выборки символов текста в функции преобразования писем в числовой вектор; длина выборки (длина «генератора» – последовательности, порождающей «ген»). **Цель исследования:** оценка влияния численного значения длины выборки (длины «генератора» – последовательности, порождающей «ген») модели электронных писем в задаче обнаружения спама с ее применением на результаты обнаружения. **Методы:** использование решающих правил импликативного типа «ЕСЛИ..., ТО...». **Результаты:** результаты эксперимента подтверждают корректность и применимость ранее разработанной авторами модели электронных писем. При этом применение модели электронных писем дает наилучшие результаты обнаружения при численном значении длины выборки (длины «генератора» – последовательности, порождающей «ген») равным 1 или 2, что подтверждается значениями агрегированной меры оценки (величины, объединяющей меры полноты и точности обнаружения и представляющей собой их среднее гармоническое). При значениях длины «генератора» больше или равных 3 результаты обнаруже-

ния ухудшается в среднем более чем на 10 %. Нецелесообразность использования значений длины «генератора» больше или равных 3 обоснована неприемлемым уменьшением значений полноты обнаружения и агрегированной меры оценки. **Практическая значимость:** предложенные авторами модель электронных писем и подход к обнаружению спама позволяют выделять последовательности символов текста, соотносящиеся с признаками определенного класса электронных писем, и сделать процесс обнаружения спама индивидуальным для получателя.

Ключевые слова: информационная безопасность, спам, обнаружение, модель электронного письма, генетический подход, генетическая модель, электронная почта, электронные почтовые сообщения, электронные письма.

S.V. Korelov[1], A.M. Petrov[1], L.Yu. Rotkov[2], A.A. Gorbunov[2]

[1]National Computer Incident Response & Coordination Center,
Moscow, Russian Federation
[2]National Research Lobachevsky State University of Nizhny Novgorod,
Nizhny Novgorod, Russian Federation

DETERMINING THE SAMPLE LENGTH OF THE ELECTRONIC LETTERS MODEL

In the conditions of intensive development of various spheres of activity of the state and society, the use of advanced information technologies becomes one of the most important and often decisive factors that determine the effectiveness of all levels of management. Information technologies hold a strong position in everyday life and in various spheres of government activity. Information systems of various types and purposes are being introduced for processing and analyzing information, forecasting and supporting management decisions making at various levels. The functioning of almost any organization in a varying degree depends on how efficiently and steadily its information system works, how reliably its information resources are protected from various information security threats, one of which is spam. It is worth noting that there have been many attempts to solve the problem of its detection once and for all. Research is ongoing in this subject area constantly. Based on its results, various approaches are proposed and implemented in practice. They are mostly based on the well-proven basic statistical processing of various parameters of emails. At the same time, the content of e-mails is not fully taken into account, which leads to an increase in the number of Type I errors. To address this issue, the authors previously proposed a model of emails that takes into account the content of emails, which often changes depending on the tasks performed by users and their changing information needs. Its key feature is that it operates with data obtained from the original text of emails and converted into a numeric vector. As the parameters of the model of emails that influence the selection of emails test segments, which are a reflection of their distinctive features, the authors identified: the number of numeric codes associated with text characters in the function of converting letters to numeric vector; the step of fetching text characters in the function of converting letters to a numeric vector; sample length (length of the «generator» - the sequence that generates the «gene»). **Purpose of the research:** to assess the influence of the sample length (length of the «generator» – the sequence that generates the «gene»), which is the email model key parameter, on results of the use of the model in spam detection. **Methods:** use of decision rules of the implicative type «IF ..., THEN ...». **Results:** the results of the experiment confirm the correctness and applicability of the email model previously developed by the authors. At the same time, the application of the e-mail model gives the best detection results when the numerical value of the sample length (the length of the «generator» – the sequence generating the «gene») is equal to 1 or 2, which is confirmed by the values of the aggregated assessment measure (the value combining the measures of completeness and accuracy of detection and representing their harmonic mean). When the length of the «generator» is greater than or equal to 3, the detection results deteriorate on average by more than 10 %. The inexpediency of using the «generator» length values greater than or equal to 3 is

justified by an unacceptable decrease in the values of the detection completeness and the aggregated assessment measure. **Practical relevance:** the authors' model of emails and an approach to spam detection allow to highlight the sequences of text characters corresponding to the characteristics of a certain class of emails, and to make the spam detection process individual for the recipient.

Keywords: information security, spam, detection, electronic letter model, genetic approach, genetic model, email, email messages, electronic letters.

Введение. В настоящее время одним из распространенных способов доставки информации и важным средством коммуникации являются электронные почтовые сообщения, обладающие многочисленными достоинствами, среди основных из которых доступность, оперативность и дешевизна при одновременных больших возможных объемах и разнообразных видах содержимого. Одним из ставших классическими бизнес-рисков, связанным с их использованием, является спам.

Существующие алгоритмы обнаружения спама в основном базируются на хорошо зарекомендовавшей себя базовой статистической обработке [например, 1–5] и не в полной мере учитывают содержание электронных писем, которая зачастую меняется в зависимости от выполняемых пользователями задач и меняющихся их информационных потребностях. А предлагаемые модели электронных сообщений в основном отражают их признаки и не учитывают содержание легальных электронных писем, что приводит к росту числа ошибок первого рода.

Очевидно, что отправители спама должны доставить свои сообщения и породить у читателей предопределенный смысл, побуждающий к определенным, необходимым отправителям, действиям. При этом любое изменение различных параметров электронных писем с целью обхода систем обнаружения спама не должно приводить к изменению этого смысла, что означает необходимость наличия определенной информации, не зависящей от наполнения письма и определяющей его смысл [6]. Основываясь на этом, авторами сделано предположение об относительной неизменности содержания спамовых писем в пределах одной смысловой рассылки (массовость рассылки подразумевает схожесть содержания электронных писем и их содержимого). Это предположение с учетом того, что похожее присуще биологическим системам, обусловило выбор генетического подхода для предложенной в [6] генетической модели электронных писем, позволяющей выделять текстовые отрезки электронных писем, являющиеся отражением их отличительных признаков:

$$\Psi_{el} = \langle gens, Gen\_Code \rangle . \quad (1)$$

Ее корректность и практическая применимость для обнаружения спама (классификации электронных писем на спамовые и легальные) продемонстрированы в [11, 12].

Ключевой особенностью данной модели является то, что она оперирует с преобразованными в числовой вектор данными, полученными из исходных текстов электронных писем. Основываясь на описанных в [7–10] положениях метода, в качестве параметров модели электронных писем, оказывающих влияние на выделение текстовых отрезков писем, являющихся отражением их отличительных признаков, выделены: q – количество числовых кодов, сопоставляемых символам текста, в функции преобразования писем в числовой вектор; Δt – шаг выборки символов текста в функции преобразования писем в числовой вектор; n – длина выборки (длина «генератора» – последовательности, порождающей «ген»).

Решение задачи понимания содержания электронных писем в целом приводит к предположению, что их смысловое значение задается комбинациями букв, цифр и знаков препинания (вместе – символов), расположенных (упорядоченных) определенном образом. Сцепление отдельных символов вместе и их расположение определенным образом придают тексту определенный смысл в целом.

Все обучающиеся алгоритмы обнаружения спама требуют подходящего для них представления электронных писем. Широкое распространение получило представление электронных писем в виде набора слов («bag of words» – мешок слов; например, [13]), появляющихся в спамовых или легальных письмах, с их количественными характеристиками. Для учета контекста появления тех или иных слов, также применяются n-граммы слов. В [14] представлен подход к обнаружению спама на основе представления текстов писем в виде n-грамм символов. Предложенный подход позволяет учитывать информацию на различных уровнях: лексическом (целые слова), слов (части слов, их части речи, число и пр.), структурном (знаки препинания). Кроме этого, он устойчив к различного рода ошибкам при написании и не зависим от языкового фактора.

Проводя аналогию между предложенной моделью и описанными подходами, авторы приходят к выводу, что по своей сути «генератор» является n-граммой символов, являющейся отражением отличительного признака электронного письма. При этом модель (1) делает возможным выделение неизменных в пределах нескольких писем участков

(именно участков, а не просто последовательностей символов и слов и их последовательностей) и позволяет снизить влияние на обнаружение спама таких условий, как, например, начальный символ (слово) письма, «мусорные» символы и слова, грамматические и пунктуационные ошибки и т.п.

Необходимо отметить, что все тексты обладают различными статистическими показателями, среди которых минимальная, максимальная и средняя длина слов в символах и предложений. Например, в русском языке средняя длина слова составляет чуть больше 5 символов [15, 16], а в английском – чуть больше 4 символов [16]. Следовательно, выбор численного значения n влияет на то, что в качестве «генератора» начинают выступать слова (предложения, группы предложений), а, следовательно, и на выделение именно неизменных участков.

Таким образом, настоящая статья посвящена исследованию вопроса влияния численного значения ключевого параметра n модели электронных писем (1) в задаче обнаружения спама с ее применением на результаты обнаружения.

1. Основные положения модели электронных писем. Для формирования модели электронных писем (1) в [6] обоснован выбор генетического подхода. При ее построении авторы исходили из следующих предположений:

1. Спамовые письма обладают специфической отличительной от легальных писем информацией, которую можно интерпретировать как «генетическую информацию» спама, закодированную определенным образом («генетическим кодом») в отдельных участках писем (отдельные последовательности символов электронных писем).

2. «Генетический код» электронных писем универсален.

3. «Генетический код» электронных писем основан на символах, составляющих электронные письма.

4. Электронные письма можно представить в виде цифровых последовательностей.

Под термином «ген» gen_j электронного письма применительно к данной области исследования понимается специальным образом выделенная последовательность символов текста, соответствующая отдельной последовательности исходных символов текста электронного письма и соотносящаяся с признаками определенного класса электронных писем.

Под «генетическим кодом» Gen_Code электронных писем понимается совокупность правил, по которым из исходной последовательности символов писем, являющихся частью текста и упорядоченных от начала к концу, выделяется конечное число соответствующих ей последовательностей [9, 10].

В качестве базового подхода к выделению значимых последовательностей использован подход к формированию математической модели текста, описанный в [7–10].

Применение модели (1) для решения задачи обнаружения спама можно описать следующими основными этапами:

1. На первом этапе происходит преобразование исходного текста электронного письма (выбраны письма в кодировке ASCII), в ходе которого каждый его символ последовательно заменяется на соответствующий ему десятичный код (один символ – одно значение кодировки). Функция преобразования имеет следующий вид:

$$Conv\_to\_Dig(el): el = (sym_0, sym_1, \ldots, sym_{M-2}, sym_{M-1}) \xrightarrow{q=256,\,\Delta t=1} el' = \\ = (b_0, b_1, \ldots, b_{M-2}, b_{M-1}), \quad (2)$$

где $el = (sym_0, sym_1, \ldots, sym_{M-2}, sym_{M-1})$ – электронное письмо, представленное в виде конечной последовательности символов; $M = |el|$ – длина электронного письма (количество символов в электронном письме); b – десятичное значение байта, соответствующего конкретному символу электронного письма по таблице ASCII.

2. На следующем этапе происходит последовательная разбивка полученной последовательности (так называемого «генетического текста») на участки разной длины («гены»). Ключевыми на этом этапе являются так называемые «генераторы» и алгоритм определения их правых границ.

Под «генератором» g участка понимается такая числовая последовательность

$$g = (b_k, b_{k+1}, \ldots, b_{k+n-2}, b_{k+n-1}), \quad (3)$$

для которой справедливы условия:

$$\begin{cases} k \in M, \\ 1 \le k \le (M - n). \end{cases} \quad (4)$$

«Генератор» участка является, по сути, его характерной единицей, обладающей следующими ключевыми параметрами: n – длина

«генератора»; k – начальная позиция «генератора» в пределах «генетического текста».

«Генератор» является параметром алгоритма определения правой границы участков, в основе работы которого лежит процедура расчета их длины, в основу которой положен алгоритм H_1 [9, 10], расчет по которому выглядит следующим образом. На первом шаге за начало участка принимается b_0. Поиск длины участка осуществляется в рамках работы цикла, ключевым действием которого является сравнение текущего кандидата в «генераторы» со всеми предыдущими в анализируемой числовой последовательности el'. При этом началом первого кандидата в «генераторы» участка является код b_1. В случае равенства текущего кандидата в «генераторы» с одним из предыдущих принимается решение о нахождении окончания текущего участка. Тем самым определяется «ген» gen_1 электронного письма с его длиной l_1, который может быть порожден «генератором» участка по начальному состоянию, расположенному в начале или внутри его. Далее с (l_1+1)-й позиции числовой последовательности el' осуществляется аналогичная процедура определения длины l_2 «гена» gen_2 и, соответственно, его границ. Процесс выделения «генов» продолжается по всей длине el'.

Таким образом, «генетический код» Gen_Code в модели электронных писем (1) описывается следующим выражением:

$$Gen\_Code = \langle Conv\_to\_Dig(el), g, H_1 \rangle. \quad (5)$$

2. Экспериментальная часть. Для проведения эксперимента были использованы два набора электронных писем:

1) англоязычных [17] (сформирован и описан в [18] с дополнительными изменениями в соответствии с [11]), состоящих из 6 групп легальных писем общим количеством 16 100 писем и 6 групп спамовых писем общим количеством 16420 писем;

2) русскоязычных, состоящих из 3 групп рассылок порталов securitylab.ru, security.nnov.ru и xakep.ru (за период с 28 апреля 2009 г. по 4 марта 2011 г.) общим количеством 1242 письма и 2 групп спамовых писем, поступивших на индивидуальные почтовые адреса двух различных адресатов одного почтового сервера в зоне .ru, общим количеством 3215 писем.

В качестве значений параметров модели электронных писем заданы следующие: $q=256$ – соответствует количеству символов кодировки Windows-1251; $\Delta t=1$ – шаг дискретизации, равен одному символу; $n=1\ldots20$.

Эксперимент и оценка его результатов проводились аналогично описанным в [11]. Для каждой категории (класса) писем (легальные и спамовые) каждой группы писем были рассчитаны наборы «генов» и определен коэффициент принадлежности каждого письма к легальным или спамовым письмам, за который принято суммарное количество «генов», содержащихся в письме, встретившихся в соответствующих категориях всех групп. Решение о принадлежности письма к спамовым или легальным принималось с использованием простейшего решающего правила – по принципу большего суммарного количества «генов» соответствующей категории. При этом для классифицируемого письма расчет набора «генов» его группы велся только для писем, стоящих перед ним в списке, что позволило частично имитировать процесс получения писем адресатом.

В качестве мер оценки результатов эксперимента использованы полнота, точность и F-мера обнаружения (классификации) [13, 19–22].

Под полнотой обнаружения спамовых и легальных писем будем понимать соотношение числа всех верно классифицированных электронных писем к числу электронных писем, которые должны были быть отнесены к тому или иному классу:

$$R=\frac{N_{corr\_a}}{N_{corr\_a}+N_{incorr\_r}}, \tag{6}$$

где N_{corr_a} – количество электронных писем, корректно отнесенных к заданной категории (истинно положительные результаты или *TP – true positive*); N_{incorr_r} – количество электронных писем, не корректно признанных не принадлежащими заданной категории (ложноотрицательные результаты или *FN – false negative*).

Иначе, полнота характеризует потери процесса классификации электронных писем. Как следует из представленной формулы, чем выше значение полноты, тем меньше потери правильных классификаций. Таким образом, R определяет способность процесса классификации электронных писем обнаруживать заданный класс вообще.

Под точностью обнаружения спамовых и легальных писем будем понимать соотношение числа верно классифицированных электронных писем к числу всех классифицированных электронных писем как принадлежащих к тому или иному классу:

$$P = \frac{N_{corr\_a}}{N_{corr\_a} + N_{incorr\_a}}, \tag{7}$$

где N_{incorr_a} – количество электронных писем, не корректно признанных принадлежащими заданной категории (ложноположительные результаты или *FP – false positive*).

Иначе точность можно интерпретировать как долю объектов, названных классификатором положительными и при этом действительно являющихся положительными. Таким образом, P определяет способность процесса классификации электронных писем правильно обнаруживать заданный класс (долю правильных классификаций), т.е. чем лучше выстроен процесс классификации, тем меньше будет неверно классифицированных электронных писем как принадлежащих заданному классу.

Безусловно, чем выше полнота и точность, тем лучше. Но очевидно, что достичь максимальную полноту и точность одновременно невозможно. В связи с этим для выбора лучшего варианта классификации использована сбалансированная *F*-мера обнаружения (классификации) [13, 19–22] спамовых и легальных писем, которая позволяет объединить полноту и точность в агрегированную величину для оценки, представляющую собой их среднее гармоническое:

$$F = \frac{2 \cdot P \cdot R}{P + R}. \tag{8}$$

Из (8) следует, что F-мера достигает максимума при полноте и точности, равных единице, и близка к нулю, если один из аргументов близок к нулю. Таким образом, F-мера позволяет определить наилучший процесс классификации электронных писем с учетом одновременно полноты и точности, т.е. чем лучше выстроен процесс классификации, тем больше значение F-меры.

Результаты эксперимента на англоязычных и русскоязычных письмах округлены до сотых долей процента по правилам простого математического округления и представлены в табл. 1 и 2 соответственно.

Таблица 1

Результаты эксперимента на наборе англоязычных электронных писем, %

Значение *n*	Полнота обнаружения *R*, %	Точность обнаружения *P*, %	Значения *F*-меры, %
1	**95,92**	96,63	**96,27**
2	**90,94**	96,44	**93,61**
3	74,75	98,00	84,81
4	62,12	98,85	76,30
5	55,70	99,35	71,38
6	49,94	98,19	66,21
7	48,03	99,04	64,69
8	46,58	99,05	63,36
9	44,91	99,08	61,81
10	42,24	99,13	59,24
11	39,92	99,02	56,90
12	38,08	98,79	54,97
13	36,67	98,86	53,50
14	35,18	98,84	51,89
15	33,91	99,18	50,54
16	32,01	97,24	48,17
17	31,42	98,92	47,69
18	30,54	99,09	46,70
19	29,85	99,13	45,89
20	29,08	99,37	44,99

Таблица 2

Результаты эксперимента на наборе русскоязычных электронных писем, %

Значение *n*	Полнота обнаружения *R*, %	Точность обнаружения *P*, %	Значения *F*-меры, %
1	**93,04**	93,74	**93,39**
2	**85,62**	91,45	**88,44**
3	69,17	90,01	78,23
4	60,98	89,97	72,69
5	59,05	96,13	73,16
6	55,01	98,08	70,49
7	54,43	100,00	70,49
8	53,06	99,96	69,32
9	42,09	99,95	59,24
10	38,75	99,88%	55,84
11	36,32	100,00	53,29

Окончание табл. 2

Значение n	Полнота обнаружения R, %	Точность обнаружения P, %	Значения F-меры, %
12	35,16	100,00	52,03
13	34,84	100,00	51,68
14	34,24	100,00	51,01
15	34,22	100,00	50,99
16	32,35	100,00	48,89
17	30,81	100,00	47,10
18	30,49	100,00	46,73
19	29,93	100,00	46,07
20	28,65	100,00	44,54

Результаты эксперимента показывают, что при значениях $n=1$ и $n=2$ значения полноты и F-меры для обоих наборов писем составляют не менее 85 %. При значениях $n\geq 3$ их значения уменьшаются в среднем более чем на 5 %, а при $n\geq 4$ и $n\geq 5$ в англоязычном и русскоязычном наборе соответственно (превышение средней длины слова в английском и русском языках) – более чем на 20 %. При этом необходимо отметить, что увеличение значения n приводит к росту точности обнаружения при одновременном неприемлемом резком уменьшении полноты обнаружения.

Выводы. Результаты эксперимента подтверждают сделанные в [11] и [12] выводы о корректности и применимости разработанной [6] авторами модели электронных писем (1), позволяющей выделять текстовые отрезки электронных писем, являющиеся отражением их отличительных признаков, для обнаружения спама. При этом предложенные авторами модели электронных писем (1) и подход к обнаружению спама позволяют выделять последовательности символов текста, соотносящиеся с признаками определенного класса электронных писем, и сделать процесс обнаружения спама индивидуальным для получателя.

Также полученные результаты подтверждают сделанные в [12] выводы о том, что применение модели электронных писем (1) дает наилучшие результаты обнаружения при численном значении ключевого параметра $n=1$ и $n=2$, что подтверждается значениями F-меры при данных значениях n. При значениях $n\geq 3$ результаты обнаружения ухудшаются в среднем более чем на 10 %. Нецелесообразность использования значений $n\geq 3$ обоснована неприемлемым уменьшением значений полноты обнаружения и F-меры.

В дальнейшем также целесообразно провести исследование вопроса влияния численного значения ключевого параметра q модели электронных писем (1) в задаче обнаружения спама с ее применением на результаты обнаружения.

Библиографический список

1. A Bayesian Approach to Filtering Junk E-Mail. In Proc. of 1998 AAAI Workshop on Learning for Text Categorization / M. Sahami, S. Dumais, D. Heckerman, E. Horvitz // AAAI Technical Report WS-98-05. – 1998. – P. 55–62.

2. Robinson G. A Statistical Approach to the Spam Problem // Linux Journal. – 2003. – Iss. 107.

3. Junejo K., Yousaf M., Karim A. A Two-Pass Statistical Approach for Automatic Personalized Spam Filtering // Proc. of ECML-PKDD Discovery Challenge Work-shop. – 2006. – P. 16–27.

4. Junejo K., Karim A. PSSF: A Novel Statistical Approach for Personalized Service-side Spam Filtering // WI-07: Proceedings of the IEEE/WIC/ACM International Conference on WebIntelligence. – 2007. – P. 228–234. DOI: https://doi.org/10.1109/WI.2007.47

5. A Memory-Based Approach to Anti-Spam Filtering for Mailing Lists / G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C.D. Spyropoulos, P. Stamatopoulos // Information Retrieval. – 2003. – Vol. 6. – P. 49–73. DOI: https://doi.org/10.1023/A:1022948414856

6. Модель электронных писем в задаче обнаружения спама / С.В. Корелов, А.М. Петров, Л.Ю. Ротков, А.А. Горбунов // Вестник Поволж. гос. технолог. ун-та. Сер. Радиотехнические и инфокоммуникационные системы. – 2020. – № 2(46). – С. 44–54. DOI: https://doi.org/10.25686/2306-2819.2020.2.44

7. Корелов С.В., Ротков Л.Ю. Метод генетических карт в задаче идентификации спама // Информационно-измерительные и управляющие системы. – 2011. – Т. 9, № 3. – С. 72.

8. Корелов С.В., Ротков Л.Ю. Идентификация текстового спама методом генетических карт // Вестник Нижегород. ун-та им. Н.И. Лобачевского. – 2012. – № 4(1). – С. 101–104.

9. Кирьянов К.Г. Генетический код и тексты: динамические и информационные модели сложных систем / ред. Л.Ю. Ротков, А.В. Якимов. – Н. Новгород: ТАЛАМ, 2002. – 100 с.

10. Кирьянов К.Г. Выбор оптимальных базовых параметров источников экспериментальных данных при их идентификации // Идентификация систем и задачи управления SICPRO'04: тр. III Междунар. конф. – М.: Изд-во ИПУ РАН, 2004. – С. 187–208.

11. К вопросу об определении численного значения параметра модели электронных писем / С.В. Корелов, А.М. Петров, Л.Ю. Ротков, А.А. Горбунов // Автоматизированные системы управления и информационные технологии (АСУИТ-2020): сб. материалов науч.-техн. конф. – Пермь: Изд-во Перм. нац. исследов. политехн. ун-та, 2020.

12. К вопросу об определении численного значения параметра в модели электронных писем / С.В. Корелов, А.М. Петров, Л.Ю. Ротков, А.А. Горбунов // Тр. XXIV науч. конф. по радиофизике, посв. 75-лет. радиофизич. фак-та (Н. Новгород, 13–31 мая 2020 г.). – Н. Новгород: Изд-во ННГУ, 2020. – С. 471–474.

13. Sebastiani F. Machine Learning in Automated Text Categorization // ACM Computing Surveys. – 2002. – Vol. 34, no. 1. – P. 1–47. DOI: https://doi.org/10.1145/505282.505283

14. Kanaris I., Kanaris K., Stamatatos E. Spam Detection Using Character N-Grams // SETN. Lecture Notes in Computer Science. – 2006. – Vol. 3955. – Springer. – P. 95–104. DOI: https://doi.org/10.1007/11752912_12

15. Шаров С.А. Частотный словарь. РосНИИ ИИ [Электронный ресурс]. – URL: http://www.artint.ru/projects/frqlist.php (дата обращения: 13.09.2020).

16. Бойков В.В., Жукова Н.А., Романова Л.А. Распределение длины слов в русских, английских и немецких текстах [Электронный ресурс]. – URL: http://tverlingua.ru/archive/001/01_1-006.htm (дата обращения: 13.09.2020).

17. Enron-Spam datasets. – URL: http://www2.aueb.gr/users/ion/data/enron-spam/

18. Metsis V., Androutsopoulos I., Paliouras G. Spam Filtering with Naive Bayes – Which Naive Bayes? // Proc. of the Third Conference on Email and Anti-Spam (CEAS 2006). – 2006. – P. 28–69.

19. Sebastiani F. Text Categorization / Alessandro Zanasi (ed.). Text Mining and its Applications. – WIT Press, Southampton, UK, 2005. – P. 109–129.

20. Aas K., Eikvil L. Text Categorisation: A Survey. – Tech. rep. 941. Norwegian Computing Center, Oslo, Norway, 1999.

21. Manning C., Raghavan P., Shütze H. Introduction to Information Retrieval. – Cambridge University Press, 2008. DOI: https://doi.org/10.1017/CBO9780511809071

22. Sokolova M., Lapalme G. A Systematic Analysis of Performance Measures for Classification Tasks // Information Processing & Management. – 2009. – Vol. 45, no. 4. – P. 427–437. DOI: https://doi.org/10.1016/j.ipm.2009.03.002

References

1. Sahami M., Dumais S., Heckerman D., Horvitz E. A Bayesian Approach to Filtering Junk E-Mail. In Proc. of 1998 AAAI Workshop on Learning for Text Categorization. *AAAI Technical Report WS-98-05*, 1998, pp. 55-62.

2. Robinson G. A Statistical Approach to the Spam Problem. *Linux Journal*, 2003, iss. 107.

3. Junejo K., Yousaf M., Karim A. A Two-Pass Statistical Approach for Automatic Personalized Spam Filtering. *Proc. of ECML-PKDD Discovery Challenge Work-shop*, 2006, pp. 16-27.

4. Junejo K., Karim A. PSSF: A Novel Statistical Approach for Personalized Service-side Spam Filtering. *WI-07: Proceedings of the IEEE/WIC/ACM International Conference on WebIntelligence*, 2007, pp. 228-234. DOI: https://doi.org/10.1109/WI.2007.47

5. Sakkis G., Androutsopoulos I., Paliouras G., Karkaletsis V., Spyropoulos C.D., Stamatopoulos P. A Memory-Based Approach to Anti-Spam Filtering for Mailing Lists. *Information Retrieval*, 2003, vol. 6, pp. 49-73. DOI: https://doi.org/10.1023/A:1022948414856

6. Korelov S.V., Petrov A.M., Rotkov L.Iu., Gorbunov A.A. Model' elektronnykh pisem v zadache obnaruzheniia spama [The Electronic Letters Model in the Spam Detection Task]. *Vestnik Povolzhskogo gosudarstvennogo tekhnologicheskogo universiteta. Radiotekhnicheskie i infokommunikatsionnye sistemy*, 2020, no. 2(46), pp. 44-54. DOI: https://doi.org/10.25686/2306-2819.2020.2.44

7. Korelov S.V., Rotkov L.Iu. Metod geneticheskikh kart v zadache identifikatsii spama [The method of genetic maps for spam detection problem]. *Informatsionno-izmeritel'nye i upravliaiushchie sistemy*, 2011, vol. 9, no. 3, 72 p.

8. Korelov S.V., Rotkov L.Iu. Identifikatsiia tekstovogo spama metodom geneticheskikh kart [Identification of text spam using genetic maps]. *Vestnik Nizhegorodckogo universiteta imeni N.I. Lobachevskogo*, 2012, no. 4(1), pp. 101-104.

9. Kir'ianov K.G. Geneticheskii kod i teksty: dinamicheskie i informatsionnye modeli slozhnykh sistem [Genetic code and texts: dynamic and information models of complex systems]. Ed. L.Iu. Rotkov, A.V. Iakimov. Nizhnii Novgorod: TALAM, 2002, 100 p.

10. Kir'ianov K.G. Vybor optimal'nykh bazovykh parametrov istochnikov eksperimental'nykh dannykh pri ikh identifikatsii [Selection of the optimal basic parameters of experimental data sources during their identification]. *Identifikatsiia sistem i zadachi upravleniia SICPRO'04. Trudy III Mezhdunarodnoi konferentsii*. Moscow: Institut problem upravleniia imeni V.A. Trapeznikova Rossiiskoi akademii nauk, 2004, pp. 187-208.

11. Korelov S.V., Petrov A.M., Rotkov L.Iu., Gorbunov A.A. K voprosu ob opredelenii chislennogo znacheniia parametra modeli elektronnykh pisem [To the Question of Determining the Numerical Value of the Electronic Letters Model Parameter]. *Avtomatizirovannye sistemy upravleniia i informatsionnye tekhnologii (ASUIT-2020). Sbornik materialov nauchno-tekhnicheskoi konferentsii*. Perm': Permskii natsional'nyi issledovatel'skii politekhnicheskii universitet, 2020.

12. Korelov S.V., Petrov A.M., Rotkov L.Iu., Gorbunov A.A. K voprosu ob opredelenii chislennogo znacheniia parametra v modeli elektronnykh pisem [To the Question of Determining the Numerical Value of the Parameter in the Electronic Letters Model]. *Trudy XXIV nauchnoi konferentsii po radiofizike, posviashchennye 75-letiiu radiofizicheskogo fakul'teta (Nizhnii Novgorod, 13-31 May 2020)*. Nizhnii Novgorod: Natsional'nyi issledovatel'skii Nizhegorodskii gosudarstvennyi universitet imeni. N.I. Lobachevskogo, 2020, pp. 471-474.

13. Sebastiani F. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 2002, vol. 34, no. 1, pp. 1-47. DOI: https://doi.org/10.1145/505282.505283

14. Kanaris I., Kanaris K., Stamatatos E. Spam Detection Using Character N-Grams. *SETN. Lecture Notes in Computer Science*, 2006, vol. 3955, Springer, pp. 95-104. DOI: https://doi.org/10.1007/11752912_12

15. Sharov S.A. Chastotnyi slovar' Rossiiskogo nauchno issledovatel'skogo instituta iskusstvennogo intellekta [Frequency dictionary.

Russian Research Institute of Artificial Intelligence], available at: http://www.artint.ru/projects/frqlist.php (accessed 13 September 2020).

16. Boikov V.V., Zhukova N.A., Romanova L.A. Raspredelenie dliny slov v russkikh, angliiskikh i nemetskikh tekstakh [Distribution of word length in Russian, English and German texts], available at: http://tverlingua.ru/archive/001/01_1-006.htm (accessed 13 September 2020).

17. Enron-Spam datasets, available at: http://www2.aueb.gr/users/ion/data/enron-spam/

18. Metsis V., Androutsopoulos I., Paliouras G. Spam Filtering with Naive Bayes – Which Naive Bayes? *Proc. of the Third Conference on Email and Anti-Spam (CEAS 2006)*, 2006, pp. 28-69.

19. Sebastiani F. Text Categorization. Alessandro Zanasi (ed.). Text Mining and its Applications. WIT Press, Southampton, UK, 2005, pp. 109-129.

20. Aas K., Eikvil L. Text Categorisation: A Survey. Tech. rep. 941. Norwegian Computing Center, Oslo, Norway, 1999.

21. Manning C., Raghavan P., Shütze H. Introduction to Information Retrieval. Cambridge University Press, 2008. DOI: https://doi.org/10.1017/CBO9780511809071

22. Sokolova M., Lapalme G. A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing & Management*, 2009, vol. 45, no. 4, pp. 427-437. DOI: https://doi.org/10.1016/j.ipm.2009.03.002

Сведения об авторах

Корелов Сергей Викторович (Москва, Россия) – сотрудник Национального координационного центра по компьютерным инцидентам (107031, Москва, ул. Б. Лубянка, 1/3, e-mail: korelovsv@cert.gov.ru).

Петров Артем Михайлович (Москва, Россия) – сотрудник Национального координационного центра по компьютерным инцидентам (107031, Москва, ул. Б. Лубянка, д. 1/3, e-mail: pam@cert.gov.ru).

Ротков Леонид Юрьевич (Нижний Новгород, Россия) – кандидат технических наук, доцент, начальник Управления информационной безопасности, заведующий кафедрой «Безопасность информационных систем Национального исследовательского Нижегородского государственного университета им. Н.И. Лобачевского (603950, Нижний Новгород, пр. Гагарина, 23, e-mail: rtv@rf.unn.ru).

Горбунов Александр Александрович (Нижний Новгород, Россия) – преподаватель кафедры «Безопасность информационных систем» Национального исследовательского Нижегородского государственного университета им. Н.И. Лобачевского (603950, Нижний Новгород, пр. Гагарина, 23, e-mail: aagor@rf.unn.ru).

About the authors

Korelov Sergei Viktorovich (Moscow, Russian Federation) is an Employee National Computer Incident Response & Coordination Center (107031, Moscow 1/3, B. Lubyanka, e-mail: korelovsv@cert.gov.ru).

Petrov Artem Mikhailovich (Moscow, Russian Federation) is an Employee National Computer Incident Response & Coordination Center (107031, Moscow 1/3, B. Lubyanka, e-mail: pam@cert.gov.ru).

Rotkov Leonid Yuryevich (Nizhny Novgorod, Russian Federation) is a Ph. D. in Technical Sciences, Associate Professor, Chief of Information Security Department, Head of Department “Security of information systems” National Research Nizhny Novgorod State University named after N.I. Lobachevsky (603950, Nizhny Novgorod, 23, pr. Gagarina, e-mail: rtv@rf.unn.ru).

Gorbunov Aleksandr Aleksandrovich (Nizhny Novgorod, Russian Federation) is a Teacher of Department “Security of information systems” National Research Nizhny Novgorod State University named after N.I. Lobachevsky (603950, Nizhny Novgorod, 23, pr. Gagarina, e-mail: aagor@rf.unn.ru).

Получено 07.10.2020