УДК 004.056, 004.852, 004.042

С.В. Корелов[1], А.М. Петров[1], Л.Ю. Ротков[2], А.А. Горбунов[2]
[1]Национальный координационный центр по компьютерным инцидентам, г. Москва
[2]Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского, Нижний Новгород

КОМБИНИРОВАНИЕ ЗНАЧЕНИЙ ПАРАМЕТРА МОДЕЛИ ЭЛЕКТРОННЫХ ПИСЕМ

В статье обсуждается вопрос комбинирования численных значений ключевого параметра n модели электронных писем для обнаружения спама, полученной на основе генетического подхода к формированию математических моделей текстов, зарекомендовавшего себя для решения различных задач.

Ключевые слова: информационная безопасность, спам, обнаружение, модель электронного письма, 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
[2]National Research Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russian Federation

COMBINING THE VALUES OF THE ELECTRONIC LETTERS MODEL PARAMETER

The article discusses the issue of combining the numerical values of key parameter n of the genetic model of emails used to detect spam, which is based on a genetic approach to the formation of texts mathematical models, proven to solve various problems.

Keywords: information security, spam, detection, electronic letter model, email, e-mail messages, electronic letters.

Введение. В настоящее время одним из широко применяемых средств коммуникации является электронная почта, которая также является обязательной для использования многих различных электронных услуг и сервисов. По оценкам The Radicati Group, Inc. [1], в 2020 году электронной почтой пользуется уже половина населения земли. При этом количество пользователей электронной почты до конца 2020 года превысит отметку в 4 млрд с прогнозом более 4,4 млрд в 2024 году.

При этом необходимо констатировать, что помимо достоинств электронная почта, к сожалению, несет и угрозы ее пользователям, среди которых одной из наиболее актуальной является спам. Так, например, спам в настоящее время является важнейшим способом доставки вредоносного программного обеспечения

и осуществления фишинговых атак [2]. Таким образом, представляется очевидным, что обнаружение спама является не просто желательной, а остронеобходимой и неотъемлемой частью общей системы безопасности информационных систем.

Необходимо отметить, что «борьба» исследователей со спамом не останавливается и идет буквально за каждые 0,01% точности и полноты обнаружения спама [3]. При этом актуальным является вопрос выбора эффективных (с точки зрения качества обнаружения спама) признаков электронных почтовых сообщений для процесса классификации, что требует проведения соответствующих исследований [3].

В связи с актуальностью и важностью данного направления исследований в задаче обнаружения спама в [4] авторами предложена и в [3] уточнена модель электронных писем, позволяющая специфическим способом выделять текстовые отрезки электронных писем, являющиеся отражением их отличительных признаков (так называемые «гены» или термы):

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

Ключевой особенностью данной модели 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 в [5-7] обоснованы:

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», N-граммы – последовательности, порождающей терм).

Там же продемонстрированы корректность и практическая применимость данной модели для обнаружения спама (классификации электронных писем на спамовые и легальные), а также обоснован выбор численного значения параметра n модели электронных писем.

Настоящая статья посвящена исследованию вопроса применения комбинированного подхода при использовании численных значений ключевого параметра n модели электронных писем (1) в задаче обнаружения спама с ее применением и его влияния на результаты обнаружения.

Основная часть. В [5-7] показано, что применение модели электронных писем (1) на англоязычном наборе электронных писем (сформирован и описан в [8] с дополнительными изменениями в соответствии с [5], состоит из 6 групп легальных писем общим количеством 16100 писем и 6 групп спамовых писем общим количеством 16420 писем) дает наилучшие результаты обнаружения при численном значении ключевого параметра $n = 1$ и $n = 2$. При значениях $n \geq 3$ полнота обнаружения ухудшается в среднем более, чем на 5%, а при $n \geq 4$

(превышение средней длины слова в английском языке [9]) – более, чем на 20%, что делает нецелесообразным использование данных значений n. При этом результаты проведенных экспериментов показывают [7], что одновременно при уменьшении полноты обнаружения увеличивается ее точность (снижается количество неверно классифицируемых писем при одновременном увеличении числа неклассифицированных писем). При этом при $n \geq 5$ значение точности практически не изменятся.

В связи с изложенным, авторами сделано предположение о возможности улучшения результатов обнаружения с применением модели (1) при использовании комбинаций значений n (в совокупности по нескольким значениям n). Для подтверждения или опровержения данной гипотезы поставлен эксперимент, в качестве значений параметров модели электронных писем для которого приняты следующие:

$q = 256$;

$\Delta t = 1$ – шаг дискретизации равен одному символу;

$n = 1$ и комбинации значений $n = 1 \div 2$; $n = 1 \div 3$; $n = 1 \div 4$; $n = 1 \div 5$.

Эксперимент и оценка его результатов проводились следующим образом.

На первом этапе для всех n для каждой категории (класса) писем (легальные и спамовые) каждой группы писем были выделены наборы термов и определены:

N_{terms} – общее количество термов в письме;

N_{terms}^{Spam} – количество термов в письме, встречающихся в спамовых письмах;

N_{terms}^{Legal} – количество термов в письме, встречающихся в легальных письмах.

Для каждого письма рассчитан коэффициент его принадлежности к классу легальных или спамовых, за который принято отношение количества термов того или иного класса к количеству термов, из которых состоит письмо:

$$K^{Legal} = \frac{N_{terms}^{Legal}}{N_{terms}}, \quad (2)$$

$$K^{Spam} = \frac{N_{terms}^{Spam}}{N_{terms}}. \quad (3)$$

Во избежание случая равенства нулю того или иного коэффициента и для обеспечения выполнимости расчетов следующего шага значения N_{terms}, N_{terms}^{Spam} и N_{terms}^{Legal} предварительно были увеличены на 1, что не оказывает влияния на соотношение коэффициентов принадлежности K^{Legal} и K^{Spam}.

На втором шаге рассчитано среднее геометрическое полученных на предыдущем этапе значений коэффициентов для заданных комбинаций значений n.

На третьем (заключительном) этапе принималось решение о принадлежности письма к спамовым или легальным с использованием простейшего решающего правила – по принципу большего значения коэффициента принадлежности. При их равных значениях письмо, как и в [5-7], считалось неклассифицированным.

Аналогично [5-7] для классифицируемого письма выделение набора термов его группы велось только для писем, стоящих перед ним в списке, что позволило частично имитировать процесс получения писем адресатом.

В качестве мер оценки результатов эксперимента использованы полнота R и точность P обнаружения (классификации) [10-14].

Результаты эксперимента представлены на рисунках 1 и 2.

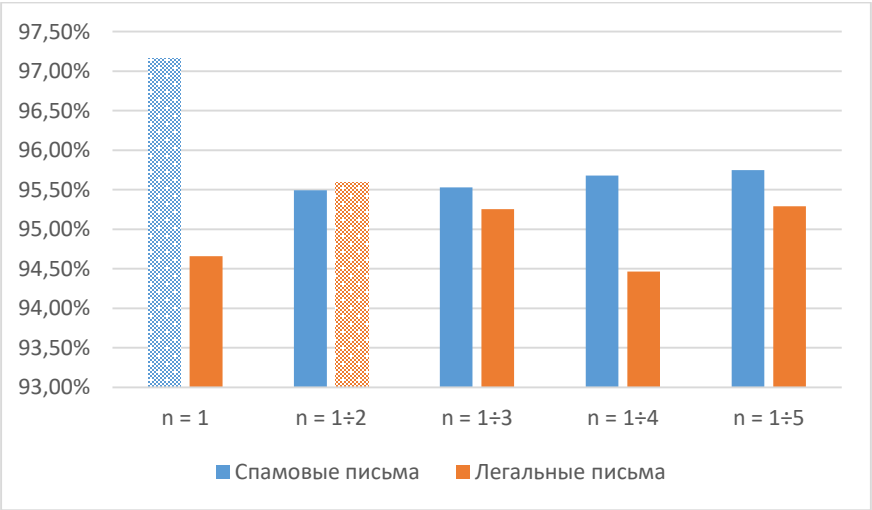

Рисунок 1 – Полнота обнаружения R в экспериментальном наборе

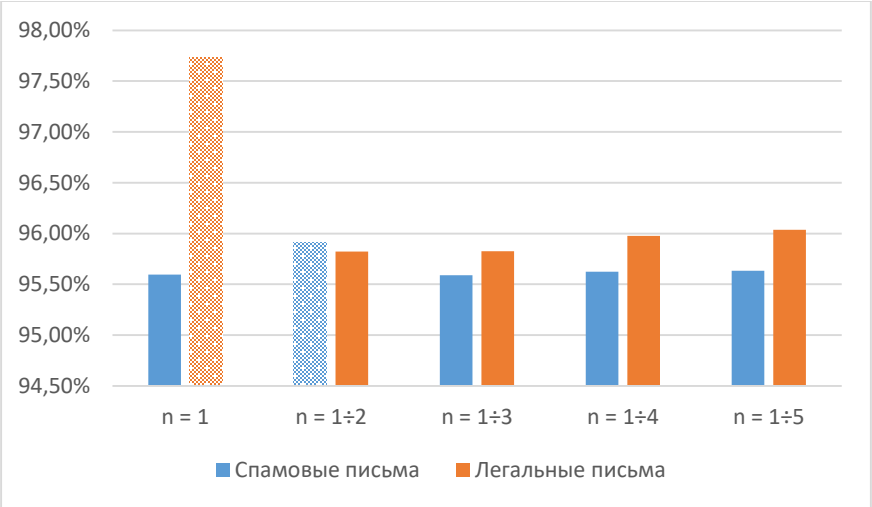

Рисунок 2 – Точность обнаружения P в экспериментальном наборе

Результаты эксперимента показывают, что при применении комбинированных подходов при использовании заданных выше комбинаций значений n полнота обнаружения спама снижается и ее значения в среднем составляют на 1,5% меньше, чем при $n = 1$. Одновременно необходимо отметить, что применение комбинированных подходов при использовании заданных комбинаций значений n позволяет увеличить полноту обнаружения легальных писем и добиться лучшей точности обнаружения спама, а следовательно снижения

количества легальных писем, неверно классифицируемых как спам. При этом наилучшее значение точности обнаружения спама достигается применением комбинированного подхода при использовании значений $n = 1 \div 2$, при котором количество неверно классифицированных легальных писем в абсолютных числах уменьшилось с 735 писем до 668 (более, чем на 9%).

Заключение. Таким образом, результаты эксперимента показывают целесообразность использования комбинированного подхода при использовании комбинаций численных значений ключевого параметра n модели электронных писем (1) в задаче обнаружения спама. Это позволяет повысить точность обнаружения спама при одновременном незначительном ухудшении полноты его обнаружения, чем, очевидно, целесообразно пренебречь в целях снижения количества легальных писем, неверно классифицируемых как спам. При этом применение модели электронных писем (1) показывает наилучшие результаты точности обнаружения спама при комбинации численных значений ключевого параметра $n = 1 \div 2$.

Список литературы

1. Email Statistics Report, 2016–2020 // The Radicati Group. URL: https://www.radicati.com/?p=13546 (дата обращения 25.11.2020).
2. Abdulhamid Sh.M., Shuaib M., Osho O., Ismaila I., Alhassan J.K. Comparative Analysis of Classification Algorithms for Email Spam Detection // International Journal of Computer Network and Information Security (IJCNIS). 2018. Vol. 10, No. 1. Pp.60-67. DOI: https://doi.org/10.5815/ijcnis.2018.01.07.
3. Корелов С.В., Петров А.М., Ротков Л.Ю., Горбунов А.А. Предобработка текстов электронных писем в задаче обнаружения спама // Труды учебных заведений связи. 2020. Т. 6. № 4. С. 80–90. DOI: https://doi.org/10.31854/1813-324X-2020-6-4-80-90.
4. Корелов С.В., Петров А.М., Ротков Л.Ю., Горбунов А.А. Модель электронных писем в задаче обнаружения спама // Вестник Поволжского государственного технологического университета. Серия: Радиотехнические и инфокоммуникационные системы. 2020. № 2 (46). С. 44-54. DOI: https://doi.org/10.25686/2306-2819.2020.2.44.
5. Корелов С.В., Петров А.М., Ротков Л.Ю., Горбунов А.А. К вопросу об определении численного значения параметра модели электронных писем // Сборник материалов научно-технической конференции «Автоматизированные системы управления и информационные технологии (АСУИТ-2020)». – г. Пермь, 2020 г. Принята к публикации 06.07.2020.
6. Корелов С.В., Петров А.М., Ротков Л.Ю., Горбунов А.А. К вопросу об определении численного значения параметра в модели электронных писем // Труды XXIV научной конференции по радиофизике, посвященной 75-летию радиофизического факультета (Нижний Новгород, 13-31 мая 2020 г.). Нижний Новгород: ННГУ, 2020. С. 471-474.

7. Корелов С.В., Петров А.М., Ротков Л.Ю., Горбунов А.А. Определение длины выборки в модели электронных писем // Вестник ПНИПУ. 2020. Принята к публикации 30.09.2020.

8. Metsis V., Androutsopoulos I., Paliouras G. Spam Filtering with Naive Bayes – Which Naive Bayes? // Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS 2006, Mountain View, USA, 27–28 July 2006). 2006. PP. 28–69.

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

10. 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.

11. Sebastiani F. Text Categorization // Zanasi A. (ed.). Text Mining and its Applications. Southampton: WIT Press, 2005. Pp. 109-129.

12. Aas K., Eikvil L. Text Categorisation: A Survey // Norwegian Computing Center. Tech. Report number: 941, 1999.

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

14. Sokolova M., Lapalme G. A Systematic Analysis of Performance Measures for Classification Tasks // Information Processing & Management. 2009. Vol. 45, Iss. 4. Pp. 427-437. DOI: https://doi.org/10.1016/j.ipm.2009.03.002.

Сведения об авторах

Корелов Сергей Викторович – сотрудник Национального координационного центра по компьютерным инцидентам, e-mail: korelovsv@cert.gov.ru

Петров Артем Михайлович – сотрудник Национального координационного центра по компьютерным инцидентам, e-mail: pam@cert.gov.ru

Ротков Леонид Юрьевич – кандидат технических наук, доцент, начальник Управления информационной безопасности, заведующий кафедрой «Безопасность информационных систем» радиофизического факультета Национального исследовательского Нижегородского государственного университета им. Н.И. Лобачевского, Нижний Новгород, email: rtv@rf.unn.ru

Горбунов Александр Александрович – преподаватель кафедры «Безопасность информационных систем» радиофизического факультета Национального исследовательского Нижегородского государственного университета им. Н.И. Лобачевского, Нижний Новгород, email: aagor@rf.unn.ru

About the authors

Sergei V. Korelov – National Computer Incident Response & Coordination Center, e-mail: korelovsv@cert.gov.ru

Artem M. Petrov – National Computer Incident Response & Coordination Center, e-mail: pam@cert.gov.ru

Leonid Yu. Rotkov – Ph. D. in Technical Sciences, Ass. Professor, Chief of Information Security Department, Head of Department «Security of information systems» of Faculty of Radiophysics, National Research Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, e-mail: rtv@rf.unn.ru

Aleksandr A. Gorbunov – Teacher of Department «Security of information systems» of Faculty of Radiophysics, National Research Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, e-mail: aagor@rf.unn.ru