УДК 621.321
DOI: 10.31673/2412-9070.2019.066199

В. В. ЗАЛИВА, аспірант;
А. П. БОНДАРЧУК, доктор техн. наук, доцент;
О. А. ЗОЛОТУХІНА, канд. техн. наук,
Державний університет телекомунікацій, Київ

ФІЛЬТРАЦІЯ СПАМУ ЕЛЕКТРОННОЇ ПОШТИ ЗА ДОПОМОГОЮ МАШИННОГО НАВЧАННЯ

У статті розглянуто фільтри для боротьби зі спамом в електронній пошті. Основну увагу приділено методам машинного навчання для успішного виявлення та фільтрація спам-листів. Охоплено важливі поняття, тенденції та оцінювання ефективності досліджень фільтрація спаму провідними постачальниками послуг електронного листування, такими як Gmail, Yahoo та Outlook. Порівнюються переваги та недоліки сторони наявних підходів щодо фільтрування спаму.

Ключові слова: машинне навчання; Spam filtering; нейронні мережі; комп'ютерна безпека; аналіз алгоритмів.

ВСТУП

Постановка проблеми. Останні п'ять років проблема спам-листів все більше привертає до себе увагу. Особу, яка надсилає спам-листи, називають спамером. Така людина може збирати електронні адреси з різних веб-сайтів, чатів та за допомогою вірусів. Спам заважає користувачеві повноцінно та ефективно використовувати свій час, ємність пам'яті та пропускну здатність мережі. Величезний обсяг спам-листів, проходячи через комп'ютерні мережі, негативно впливає на простір пам'яті серверів електронної пошти, пропускну здатність зв'язку, потужність процесорів та час користування. Загроза спаму в електронній пошті збільшується щорічно і відповідає за понад 77% усього глобального трафіку електронної пошти. Користувачі, які отримують спам-листи, залишаються невдоволеними якістю послуг сервісу електронної пошти, що може призвести до фінансових втрат компаній, які надають послуги електронного листування. Також користувачі електронної пошти можуть стати жертвами різноманітних інтернет-афер та інших шахрайських дій спамерів, які надсилають електронні листи, видаючи себе за авторитетні компанії, з метою переконати користувача розкрити конфіденційну інформацію, таку як паролі, номери кредитних карток та реєстраційні номери банку (BVN). Актуальність даної статті полягає в огляді сучасних методів машинного навчання для боротьби зі спамом та фішинг-листами.

Аналіз основних досліджень і публікацій. Методам машинного навчання останніми роками присвячено багато наукових праць, серед яких, зокрема, можна виокремити роботи таких авторів, як Гришко А. О., Бродкевич В. М., Лавренюк М. С, Новіков О. М.

ОСНОВНА ЧАСТИНА

Огляд поширення та проблем спам-листів

Використання електронної пошти в усьому світі продовжує зростати дуже швидкими темпами. У 2015 році кількість користувачів електронної пошти по всьому світу налічувалось майже 2,6 млрд. До кінця 2019 року вона примножиться до понад 2,9 млрд, тобто більше однієї третини населення світу буде використовувати електронну пошту. Хоча й існує ширше використання соціальних мереж та інших форм спілкування, електронна пошта продовжує демонструвати стабільне зростання, оскільки всі чати, соціальні мережі та інші служби вимагають від користувачів їхньої електронної адреси для доступу до своїх послуг. Окрім того, для всіх онлайн-транзакцій (наприклад, закупівель, банківських послуг тощо) потрібна дійсна електронна адреса. Оцінки (Statista, 2019) полягають у тому, що трохи менш за 60% вхідного ділового електронного трафіку — це небажана групова електронна пошта (відома як спам), яка була на найнижчому рівні з 2003 року. Однак навіть якщо глобальний відсоток спаму/не-спаму зменшується, конкуренція між спамерами та методами фільтрації спаму беззупинна. Варто зауважити, що проблема не усувається, а потреба у надійних фільтрах проти спаму залишається високою. Ідея автоматичної класифікації спам/не-спам електронних листів за допомогою методів машинного навчання була доволі популярною в наукових колах і залишається сьогодні цікавою для багатьох дослідників.

Далі пропонуємо статистику щодо пропорцій спам-листів до звичайних електронних листів (рис. 1), а також статистику стосовно зміни кількості спам-листів (у відсотках) за місяцями у рік.

Щоб ефективно боротися із загрозою, яку створюють спам-листи, провідні постачальники послуг електронної пошти, такі як Gmail, Yahoo та Outlook, застосовували в своїх спам-фільтрах комбінації з різних методів машинного навчання (ML), зокрема Neural Networks. Методи ML мають можливість вивчити та ідентифікувати спам-повідомлення та фішинг-повідомлення, аналізуючи навантаження таких повідомлень у величезній колекції комп'ютерів. Оскільки машинне

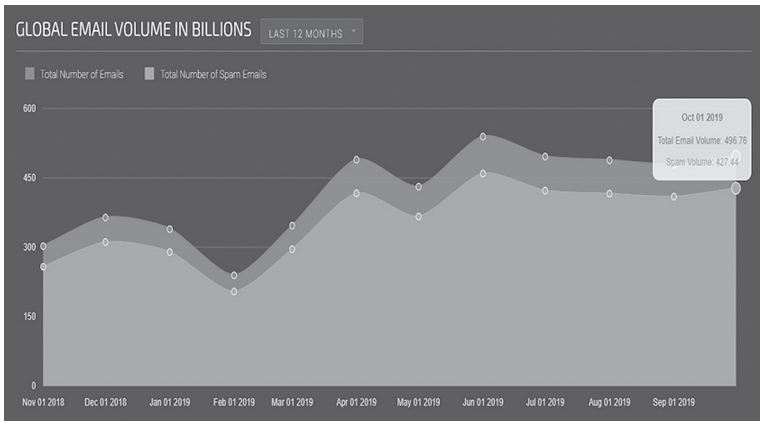


Рис. 1. Відношення поштових листів до спам-листів у 2019 році [1]

DATE	AVERAGE DAILY SPAM VOLUME (BILLIONS)	SPAM VOLUME CHANGE
2018 November	257.75	-11% ↓
2018 December	311.24	+21% ↑
2019 January	289.71	-7% ↓
2019 February	204.19	-30% ↓
2019 March	295.67	+45% ↑
2019 April	416.78	+41% ↑
2019 May	366.51	-12% ↓
2019 June	459.40	+25% ↑
2019 July	422.49	-8% ↓
2019 August	416.04	-2% ↓
2019 September	409.51	-2% ↓
2019 October	427.44	+4% ↑

Рис. 2. Зміна обсягу спам-листів із 2018 по 2019 рік [1]

навчання має можливість адаптуватися до різних умов, поштові фільтри від Gmail і Yahoo виконують більше завдань, ніж просто перевіряють непотрібні електронні листи за допомогою існуючих правил. Вони самі, фільтруючи спам-листи, генерують нові правила на основі того, про що вони дізналися. Модель машинного навчання, яка використовується у Google, вже вдосконалилася до тих меж, що може виявляти та фільтрувати спам та фішинг-листи з майже 99,9-відсотковою точністю. Наслідком цього є те, що одному з тисячі повідомлень вдалося ухилитися від фільтра спаму електронної пошти. Статистика від Google підтвердила, що 50-70% електронних листів, які отримує Gmail, є небажаною поштою. Моделі виявлення Google також містили в собі інструменти під назвою «Безпечний перегляд Google» для розпізнавання веб-сайтів зі шкідливими URL-адресами. Ефективність виявлення фішингу в Google було покращено завдяки впровадженню системи, яка затримує доставлення деяких повідомлень Gmail на деякий час, аби виконати додаткове та всебічне вивчення фішинг-повідомлень, оскільки їх легше виявити, коли вони аналізуються разом. Ця навмисна затримка впливає лише на 0,05% електронних листів.

Опис та порівняння алгоритмів фільтрації спаму

Інженерія знань та машинне навчання — це два основних підходи, які вчені застосували для подолання проблеми фільтрації спаму. Перше вирішення зосереджується на створенні системи, заснованої на знаннях, в якій заздалегідь визначено правила. Основним недоліком цього методу є те, що ці правила потрібно постійно підтримувати та оновлювати користувачем або третьою стороною, як, наприклад, постачальником програмного забезпечення. Підхід до машинного навчання, навпаки, не вимагає заздалегідь визначених правил, а лише потребує навчального набору даних, який використовуватиметься для адаптації алгоритму до моделі. Можна сказати, що алгоритм відокремлює правила класифікації від даних тесту. У цьому дослідженні порівнюються три алгоритми, придатні для завдань класифікації. Зокрема, розглядаються такі методи:

- випадковий ліс;
- k-найближчі сусіди;
- метод опорних методів із лінійним ядром.

♦ ***Випадковий ліс***. Алгоритм виводить позначку класифікації нових документів із набору дерев розв'язків, де для кожного дерева здійснюється вибірка з навчальних даних, а дерево розв'язків

створюється вибором випадкової підмножини всіх функцій (рис. 3). Алгоритм може бути застосованим для складних завдань класифікації з малим набором даних. Завдяки усередненню кількох дерев, моделі на основі випадкового лісу мають значно нижчий ризик перевитрат та меншу дисперсію порівняно з деревами прийняття рішень. Основним недоліком є продуктивність, оскільки чимала кількість дерев може уповільнити метод для прогнозування в реальному часі.

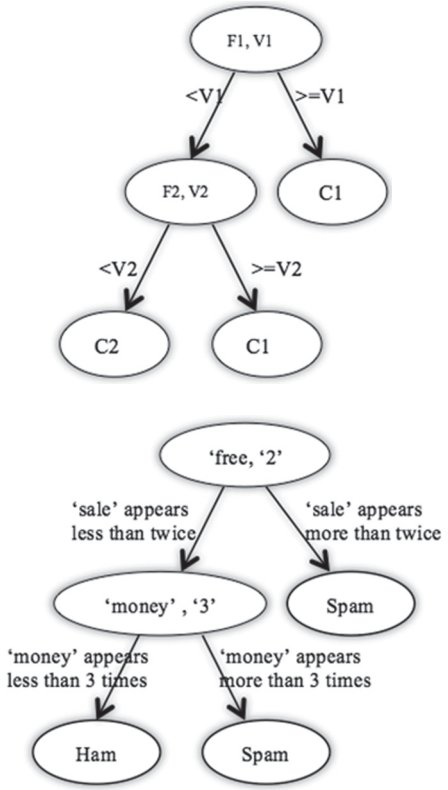


Рис. 3. Алгоритм дерева розв'язку для фільтрації спаму по електронній пошті [2]

Приклад алгоритму випадкового лісу для розв'язання проблем фільтрації спаму в електронних листах:

1: Input X: кількість вузлів

2: Input N: кількість функцій електронного повідомлення

3: Input Y: кількість дерев, які потрібно виростити

4: Поки умови не відповідають дійсності

5: Виберіть тренувальний електронний лист S у навчальному корпусі Y

6: Створіть дерево R?? із вибраного тренувального повідомлення S

7: Виберіть n ознак довільно з N; де n << N

8: Обчисліть оптимальну точку ділення для вузла d серед n функцій

9: Розділіть батьківський вузол на два дочірні вузли через оптимальне ділення

10: Виконайте кроки 1–3 до створення максимальної кількості вузлів (x)

11: Створіть свій ліс, повторивши кроки 1–4 протягом Y кількості разів

12: end while

13: Генерувати результат для кожного створеного дерева {Rt} 1Y

14: Використовувати нове повідомлення електронної пошти для кожного створеного дерева, що починається з кореневого вузла

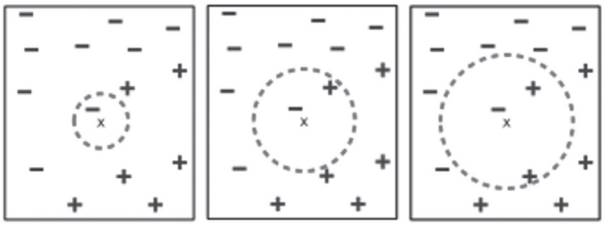
15: Позначте повідомлення електронної пошти для групи, сумісної з вузлом

16: Об'єднайте результати кожного дерева

17: return остаточну класифікацію повідомлень електронної пошти (спам/не-спам) у групу, яка має найбільше голосів (G)

18: end

♦ ***k-Найближчі сусіди.*** Класифікатор *k*-найближчого сусіда (*kNN*) — це простий метод, що добре працює з простими проблемами розпізнавання (рис. 4). Він уважається класифікатором на основі прикладу, оскільки навчальні дані використовуються для порівняння, а не для явного подання категорій. У літературі термін ледачий-учень також часто пов'язано з *kNN*. Якщо новий документ потрібно класифікувати, то *kNN* намагається знайти *k*-найближчих сусідів (більшість подібних документів) у навчальному наборі даних. З огляду на те, що знайдено та класифіковано достатньо сусідів, *kNN* використовує свій профіль, щоб призначити новий документ до тієї самої категорії. Це порівняння є процесом у реальному часі, і тому головним недоліком такого підходу є те, що алгоритм *kNN* має обчислити відстань та відсортувати всі навчальні дані для кожного прогнозування, що може бути повільним, якщо отримати великий набір даних про навчання.



а *б* *в*

Рис. 4. *k*-Алгоритм найближчих сусідів для $k = 1$ (*а*); $k = 2$ (*б*); $k = 3$ (*в*) [2]

Приклад алгоритму *kNN* для розв'язання проблем фільтрації спаму в електронних листах:

1: Знайти мітки класів повідомлення електронною поштою

2: Input k, кількість найближчих сусідів

3: Input D, набір тестового повідомлення електронної пошти

4: Input T, набір навчальних повідомлень електронної пошти

5: L-набір позначок тестового повідомлення електронної пошти

6: Read DataFile (TrainingData)
7: Read DataFile (TestingData)
8: for each d in D and each t in T do
9: Neighbors(d) = {}
10: if |Neighbors (d) | < k then
11: Neighbors(d) = Closest (d, t) Neighbors(d)
12: end if
13: if |Neighbors(d)| ≥ k then
14: restrain(M, xj, yj)
15: end if
16: end for 17: повернення остаточної класифікації повідомлення електронної пошти (спам/дійсна електронна пошта)
18: end

♦ ***Метод опорних методів із лінійним ядром.*** Оригінальний алгоритм *Support Vector Machines* (SVM) було розроблено Володимиром Вапником та Олексієм Червоненкісом у 1963 році. SVM має свою основу в широкій концепції площин прийняття рішень, які визначають межі прийняття рішень. Площини розв'язків відокремлюють поодинокі об'єкти, відшукуючи оптимальну гіперплану з максимальним запасом між двома окремими класами (рис. 5). SVM забезпечує високу точність на невеликих наборах даних, але, як правило, менш ефективно алгоритм працює з більш великими наборами даних.

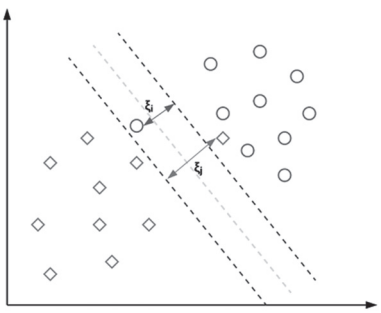


Рис. 5. Техніка формування запасів SVM у дії [3]

Приклад алгоритму SVM для розв'язання проблем фільтрації спаму в електронних листах:

1: Input тестове повідомлення електронної пошти x для класифікації
2: Навчальний набір S, функція ядра, {c1, c2,... cnum} і {,,...}
3: Кількість найближчих сусідів k
4: for i = 1 to num
5: set C=Ci;
6: for j = 1 to q
7: set =;
8: Вивести навчений класифікатор SVM f (x) через параметр склеювання (C,);
9: if (f (x) перша вироблена дискримінантна функція), then
10: keep f (x) як найбільш ідеальний класифікатор SVM (x);
11: else
12: Порівняйте класифікатор f (x) та поточний найкращий класифікатор SVM (x), використовуючи k-кратну перехресну перевірку
13: Зберігайте класифікатор з кращою точністю
14: end if
15: end for
16: end for
17: return остаточну класифікацію повідомлення електронною поштою (спам / не-спам)
18: end

Аналіз ефективності алгоритмів

Згідно з розглянутими алгоритмами було виконано тестову модель навчання кожного типу алгоритму. Порівняння ефективності всіх трьох підходів було оцінено за найчастіше використовуваними показниками: точність спаму (SP), відклик спаму (SR) та точність (A) (рис. 6). Усі три показники походять із матриці плутанини кожної моделі. Більш детально про показники:

- точність спаму (SP) — відсоток правильних результатів, поділений на кількість усіх повернених результатів;
- відклик спаму (SR) — відсоток усіх спам-листів, що правильно класифікуються як спам;
- точність (A) — відсоток усіх правильно класифікованих електронних листів.

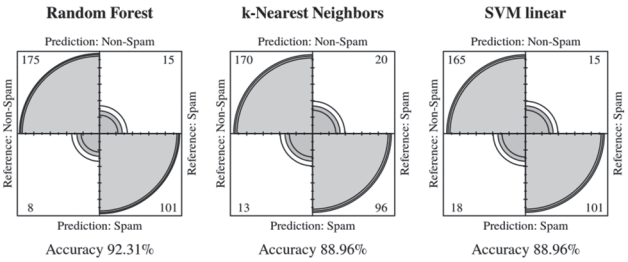


Рис. 6. Тестування трьох найголовніших алгоритмів для тестування спаму [4]

Результати роботи всіх трьох методів машинного навчання наведено на рис. 7. Із результатів випливає, що *k*-найближчі сусіди (kNN) та метод опорних методів із лінійним ядром (SVM) мають слабкі показники щодо точності, а випадковий ліс (RF) перевершує обидва інші алгоритма. Отже, RF та SVM мають порівняно однаково високий відсоток відклику спаму, тоді як kNN у цій категорії значно гірший, RF має найвищий відсоток точності спаму, а SVM — майже на 10 балів менше.

Algorithm	Spam Precision (SP)	Spam Recall (SR)	Accuracy (A)
Random Forest	92.66	87.07	92.31
k-Nearest Neighbours	88.07	82.76	88.96
SVM Linear	94.87	87.07	88.96

Рис. 7. Результати тестування алгоритмів на SP, SR та A [4]

ВИСНОВКИ

У статті було розглянуто підходи машинного навчання та їх застосування у сфері фільтрації

спаму, а також етапи еволюції спам-повідомлень. Проаналізовано основну архітектуру спам-фільтрів електронної пошти та процеси, що беруть участь у фільтрації спам-листів. У статті досліджено деякі загальнодоступні набори даних та показники ефективності, якими можна користуватись для оцінювання будь-якого спам-фільтра. Хоч дослідники і докладають зусиль щодо підвищення прогностичної точності фільтра, спамери також розвиваються і намагаються перевершити ефективність спам-фільтрів. Дуже важливим є розв'язання завдання щодо розроблення більш ефективних методів, які адекватно впораються з тенденцією чи прогресуванням функцій спаму. Також було проведено аналіз трьох найважливіших алгоритмів для фільтрації спаму. Дослідження алгоритмів показало, що найвищу ефективність згідно з параметрами SP, SR, A має алгоритм випадкового лісу.

Список використаної літератури

1. ***Total*** *Global Email & Spam Volume for October 2019 [Електронний ресурс]. URL:*

https://talosintelligence.com/reputation_center/email_rep (дата звернення: 04.11.2019).

2. ***Email*** *Spam Detection Using Mashine Learning [Електронний ресурс]. URL:*

https://ese.wustl.edu/ContentFiles/Research/UndergraduateResearch/CompletedProjects/WebPages/sp14/SongSteimle/WebPage/classifiers.html (дата звернення: 04.11.2019).

3. ***Support*** *Vector Machines — Soft Margin Formulation and Kernel Trick [Електронний ресурс]. URL:*

https://towardsdatascience.com/support-vector-machines-soft-margin-formulation-and-kernel-trick-4c9729dc8efe (дата звернення: 04.11.2019).

4. ***Comparison*** *of machine learning techniques in email spam detection [Електронний ресурс]. URL:*

https://dev.to/matchilling/comparison-of-machine-learning-techniques-in-email-spam-detection-2p0h#fn3 (дата звернення: 04.11.2019).

5. ***Albelwi S., Mahmood A.*** *A framework for designing the architectures of deep convolutional neural networks // Entropy. 2017. 19 (6). P. 242 [Електронний ресурс]. URL:*

https://www.mdpi.com/1099-4300/19/6/242 (дата звернення: 04.11.2019).

6. ***Sharma A., Suryawansi A.*** *A novel method for detecting spam email using KNN classification with spearman correlation as distance measure // Int. J. Comput. Appl. 2016. 136 (6). P. 28–34 [Електронний ресурс]. URL:*

https://pdfs.semanticscholar.org/3f1c/20b2c3b28a0328bfc5db19b02621e5874cee.pdf (дата звернення: 04.11.2019).

7. ***Deng L., Deep D. Yu.*** *Learning: Methods and Applications Now publishers. Boston, 2014 [Електронний ресурс]. URL:*

https://www.nowpublishers.com/article/Details/SIG-039 (дата звернення: 04.11.2019).

8. ***Машинне*** *навчання простими словами [Електронний ресурс]. URL:*

http://www.mmf.lnu.edu.ua/ar/1739 (дата звернення: 04.11.2019).

9. ***Akshita Tyagi.*** *Content Based Spam Classification- A Deep Learning Approach A Thesis Submitted To The Faculty Of Graduate Studies University Of Calgary. Alberta, Canada, 2016 [Електронний ресурс]. URL:*

https://prism.ucalgary.ca/handle/11023/3478 (дата звернення: 04.10.2019).

В. В. Залива, А. П. Бондарчук, О. А. Золотухина

ФИЛЬТРАЦИЯ СПАМА ЭЛЕКТРОННОЙ ПОЧТЫ С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ

В данной статье рассматриваются фильтры для борьбы против спама по электронной почте. Основное внимание уделяется методам машинного обучения для успешного выявления и фильтрации спам-писем. Статья охватывает важные понятия, тенденции и оценку эффективности исследований фильтрации спама ведущими поставщиками услуг электронной почты, такими как Gmail, Yahoo и Outlook. В статье сравниваются сильные и слабые стороны существующих подходов для фильтрации спама. В статье были рассмотрены этапы эволюции спам-сообщений. Так же особое внимание было уделено процессам, которые участвуют в фильтровании спама. В статье, был проведен анализ трех основных алгоритмов для фильтрации спама. Исследование показало, что наивысшую эффективность относительно параметров SP, SR, A показал алгоритм случайного леса. Были подняты проблемы развития борьбы со спам-сообщениями, а также, проблемы связаны с проведением различных исследований в этой области. В статье была приведена статистика количества спам-сообщений, за период с 2018 по 2019 год. Согласно статистике, за 2019 год количество спам-сообщений продолжает возрастать и достигает отметки в 500 миллиардов сообщений в месяц. Опираясь на эти данные, с уверенностью можно сказать что актуальность проблемы спама не только не утрачивает свою актуальность с годами, а и набирает ее. Ежегодно количество спам сообщений увеличивается на 20 – 40% и составляет около 77% всего почтового трафика. Спам мешает пользователям эффективно использовать инструменты электронной почты. Наибольший ущерб пользователю приносят фишинг-системы, которые провоцируют пользователей оставлять на вредоносных сайтах свои контактные данные, данные кредитных карт и паспортные данные. Алгоритмы для фильтрации спама, которые были приведены в статье, продолжают ежедневно развиваться, дополняются новыми вводными данными, чему помогает машинное обучение алгоритмов, и решают все более острые и глобальные проблемы в направлении борьбы со спамом.

Ключевые слова: машинное обучение; Spam filtering; нейронные сети; компьютерная безопасность; анализ алгоритмов.

V. V. Zaliva, A. P. Bondarchuk, O. A. Zolotukhina

E-MAIL SPAM FILTERING BY MACHINE LEARNING

This article discusses email anti-spam filters. The focus is on machine learning methods for successful detection and filtering of spam e-mails. The article covers important concepts, trends, and evaluations of the effectiveness of spam filtering research by leading e-mail providers such as Gmail, Yahoo, and Outlook. The article compares the strengths and weaknesses of existing approaches to spam filtering. The article deals with the stages of evolution of spam-messages. As the special attention has been given to processes which participate in filtering of a spam. In article, the analysis of three basic algorithms for a filtration of a spam has been spent. The research has shown that the highest efficiency concerning SP, SR, A parameters has shown the algorithm of a random forest. Problems in the development of the fight against spam-messages were raised, as well as problems associated with various studies in this area. The article provided statistics on the number of spam messages for the period from 2018 to 2019. According to the statistics, in 2019 the number of spam messages continues to increase and reaches the mark of 500 billion messages per month. Based on these data, it is safe to say that the relevance of the problem of spam not only does not lose its relevance over the years, and gaining it. Annually, the number of spam messages increases by 20 – 40% and is about 77% of all mail traffic. Spam prevents users from using e-mail tools effectively. Fishing, which is the system that provokes users to leave their contact, credit card and passport information on malicious sites, is the most damaging to the user. The spam filtering algorithms discussed in the article continue to evolve on a daily basis, with new input data being added to help machine learning of the algorithms, and to solve increasingly acute and global problems in the direction of combating spam.

Keywords: machine learning; Spam filtering; neural networks; computer security; algorithm analysis.

ЗВ'ЯЗОК

Наукове видання

Редакційна обробка та коректура
Т. В. Ількевич

Комп'ютерна верстка та дизайн
Г. С. Тимченко

Відповідальний за випуск
І. І. Тищенко

Формат 60×84/8. Папір друкарський.
Гарнітура SchoolBookC, EuropeCond. Зам. 73
Наклад 300 прим.

Державний університет телекомунікацій
03110, м. Київ, вул. Солом'янська, 7
Тел. (044) 249-25-75
E-mail: *zviaz-ok@ukr.net*