УДК 681.3.06

МЕТОДОЛОГІЯ КЛАСИФІКАЦІЇ ЛИСТІВ ЕЛЕКТРОННОЇ ПОШТИ З ВИКОРИСТАННЯМ НЕЙРОННИХ МЕРЕЖ

Ігор Терейковський

Ефективність захисту електронної пошти багато в чому залежить від достовірності розпізнавання в листах спаму та витоків. Існуючі засоби розпізнавання базуються на статистичних методах аналізу текстової інформації, що значно обмежує їх можливості щодо виявлення нових видів спаму та витоків. З метою подолання цього недоліку запропоновано методологію класифікації листів на основі змістовного аналізу електронного за допомогою нейронних мереж. Запропоновано використовувати в якості вхідних параметрів нейронної мережі частоту зустрічі в тексті листа інформативних слів в канонічні формі. Доведено, що оптимальним типом нейромережевої моделі є карта Кохонена, основною перевагою якої є висока швидкість навчання та можливість зручної візуалізації результатів класифікації. Це дозволяє швидко реагувати на нові вити спаму та витоків та проводити остаточну класифікацію листів самим користувачем. Проведені експерименти підтвердили ефективність запропонованої методології.

Ключові слова: *електронна пошта, спам, витік, нейронна мережа, карта Кохонена.*

Вступ

В теперішній час інформаційна безпека системи електронного документообігу багато в чому залежить від ефективності захисту електронної пошти від спаму та від витоків інформації. Слід зазначити, що традиційно для захисту від спаму використовуються антиспамові системи, а для захисту від витоків – DLP-системи. Принципово функціонування таких систем захисту полягає в виявленні та наступному блокуванні небезпечних електронних листів. В антиспамових системах до небезпечних листів відносять спам, а в DLP-системах – листи які містять заборонену для передачі інформацію. В більшості систем захисту для виявлення небезпечних листів використовуються методи класифікації електронної пошти на основі:

– аналізу певних формальних ознак листа (адреси відправника/отримувача, обсягу тексту, терміну відправлення);

– пошуку в тексті листа визначених слів, словосполучень та регулярних виразів.

При цьому практично не враховується зміст тексту перевіряємого листа, що разом з багатоваріантністю передачі тексту на природній мові призводить до низької точності класифікації, а як наслідок і до низької ефективності антиспамових та DLP-системи в цілому. Перспективним шляхом вирішення цієї проблеми може стати розробка методології змістовної класифікації на базі нейронних мереж, що вже довели свою ефективність в подібних системах призначених для опрацювання природньомовної інформації [2, 3].

Принципи класифікації

Базуючись на підходах по нейромережевого аналізу природньомовних текстів [1, 2] встановлено, що методологія класифікації листів електронної пошти з використанням нейронних мереж складається із наступних етапів:

– формування номенклатури вхідних параметрів;

– формування номенклатури вихідних параметрів;

– визначення оптимального типу нейромережевої моделі;

– розрахунок параметрів нейромережевої моделі обраного типу;

– верифікація моделі та уточнення її параметрів.

Також в [3, 4] доведено, що в якості вхідних параметрів нейромережевої моделі будуть використані частоти зустрічі в тексті інформативних слів в канонічній формі. В даному випадку під поняттям канонічної форми слова слід розуміти запис слова в такому вигляді, який дозволяє формувати будь-яку із його словоформ. Причиною застосування канонічних форм є те, що в українській та російській мовах на які орієнтована методологія класифікації, більшість слів можуть бути представлені в декількох словоформах без зміни своєї інформативності. Для отримання канонічних форм слів застосована методика [1, 2], яка передбачає використання словників словоформ.

Базуючись на [4] встановлено, що вихід нейромережевої моделі як мінімум повинен сигналізувати про ймовірність віднесення тексту електронного листа до класу небезпечних, підозрілих або безпечних листів. Крім того, бажано класифікувати зміст листів по тематикам.

Визначення оптимального типу нейромережевої моделі реалізовано на основі методу оптимізації, наведеному в [5]. Метод базується на підході – із множини допустимих оптимальним є

той тип нейромережевої моделі характеристики якої найбільш повно відповідають умовам поставленої задачі захисту інформації. Зазначено, що допустима множина складається таких типів нейромережевих моделей, як багатошаровий перспетрон (БШП), мережа радіальної базисної функції (РБФ), ймовірнісна мережа (PNN), мережа адаптивної резонансної теорії (АРТ), топографічна карта Кохонена (ТК), семантична нейронна мережа (СНМ), асоціативні нейронні мережі (АНМ). Таким чином множина допустимих типів нейромережевих моделей (A) має наступний вигляд

$$A=\{БШП, РБФ, PNN, ТК, АРТ, АНМ, СНМ\}. \quad (1)$$

З врахуванням (1), в загальному вигляді постановку оптимізаційної задачі можна записати так:

$$E(a_i)\to \max,\ a_i \in A,\ i=1,2,...,7, \quad (2)$$

де E – множина критеріїв оптимізації, a_i – і-ий тип нейромережевої моделі.

При цьому

$$E=\{E_k\}_K, \quad (3)$$

де E_k – k-ий критерій оптимізації, K – кількість критеріїв оптимізації.

Будь-який k-ий критерій оптимізації визначає наскільки k-та умова поставленої задачі може бути забезпечена в нейромережевій моделі. Оцінювати забезпеченість пропонується по трьохбальній шкалі: 1 – умова забезпечується повністю, -1 – умова не забезпечується, 0 – забезпечується частково.

Відповідно постановку оптимізаційної задачі (2) можна уточнити так:

$$E_\Sigma(a_i)=\sum_{k=1}^{K} E_k(a_i)\to \max,\ a_i \in A,\ i=1,2,...,7, \quad (4)$$

де E_Σ – інтегральний критерій оптимізації, K – кількість значущих критеріїв.

Для конкретної задачі захисту кількість значущих критеріїв та величину кожного з них слід уточнити. Уточнення можна реалізувати, наприклад, методом експертних оцінок. Врахувати думку експертів можливо ввівши в (4) відповідні вагові коефіцієнти. При цьому перелік основних умов задач захисту інформації які принципово можуть бути забезпечені в нейронній мережі та оцінки відповідності типу нейромережевої моделі до кожної із умов наведено в [5].

Таким чином, для визначення оптимального типу нейромережевої моделі слід:

– визначити умови поставленої задачі;

– розрахувати одиничні критерії оптимізації для кожного допустимого типу нейромережевої моделі;

– розрахувати інтегральний критерій оптимізації для кожного із допустимих типів нейромережевої моделі;

– визначити тип нейромережевої моделі з максимальною величиною інтегрального критерію.

Базуючись на результатах [1, 3, 5] сформовано перелік умов поставлених задач розпізнавання спам-листів електронної пошти та витоків інформації, які повинні забезпечуватись в нейромережевій моделі:

– в загальному випадку кількість вхідних параметрів (частоти інформативних слів в канонічній формі) обмежена кількістю інформативних слів, що використовуються в природній мові на яку орієнтована система розпізнавання. Для української та російської мов це більше ніж 10000 слів. З позицій технічної реалізації така кількість вхідних параметрів відповідає нейромережевим моделям з необмеженою кількістю вхідних параметрів. Однак в багатьох випадках тематика безпечних, підозрілих та заборонених текстів досить обмежена, що визначає і обмеженість інформативних слів які входять до їх складу. В результаті аналізу зібраних автором статистичних даних виявлено, що приблизна кількість таких слів, а значить і кількість вхідних параметрів – 1000. Таким чином не є обов'язковим виконання вимоги використання навчальних прикладів з необмеженою кількістю вхідних параметрів;

– кількість навчальних прикладів принципово не обмежена;

– в навчальних прикладах допустимі помилки;

– навчальні приклади можуть бути корельовані між собою;

– в навчальній вибірці неможливо відобразити всі можливі комбінації інформативних слів для різних типів текстів;

– в навчальній вибірці можливо пропорційно представити приклади, що відповідають безпечним, підозрілим та забороненим текстам;

– доцільно використання неперервних вхідних параметрів;

– в багатьох випадках обсяг навчальної вибірки може бути меншим від кількості вхідних параметрів;

– нейронна мережа повинна враховувати специфіку конкретного користувача, а значить не

може бути навчена в лабораторних умовах. Тому одним із найбільш важливих критеріїв оптимізації є можливість короткого терміну навчання;

– в навчальних прикладах можливо представити очікуваний вихідний сигнал (безпечний/підозрілий/заборонений текст), що дозволяє використати навчання "з вчителем";

– процес навчання повинен бути максимально автоматизованим;

– НМ повинна оперативно реагувати на нові види тексту не представлені в навчальній вибірці, тому обов'язковою є можливість донавчання НМ. Слід зазначити, що альтернативним шляхом оперативного реагування є повне перенавчання НМ. Такий підхід можна реалізувати для НМ з коротким терміном навчання;

– оскільки в системі класифікації передбачено використання буферного класу підозрілих текстів, а одиничні факти неправильної класифікації як в задачі розпізнавання спаму так і в задачі розпізнавання витоків не призводять до катастрофічних наслідків, то вимоги до якості навчання НМ можуть бути не надто високими;

– виходячи з того, що в більшості систем захисту від спаму та в DLP-системам очікувана кількість навчальних прикладів знаходиться в межах 10000, вимоги до обсягу пам'яті НМ не високі;

– оскільки в навчальній вибірці можливо представити практично всі типи безпечних текстів, то забезпечення екстраполяції результатів навчання за межі навчальної вибірки є бажаною, але не обов'язковою умовою;

– результати навчання мають бути незмінними;

– відповідно запропонованому методу класифікації слід забезпечити можливість інтерпретації вихідної інформації НМ у вигляді ймовірності;

– оскільки в сучасних системах якість автоматичної класифікації текстової інформації не повністю задовольняє практичним вимогам, то доцільна інтерпретація виходу НМ у графічному вигляді. За рахунок цього користувач системи отримає можливість автоматизованої класифікації листів;

– вербалізація НМ не обов'язкова;

– навчена НМ повинна максимально швидко проводити класифікацію;

– обсяг програмної реалізації не має суттєвого значення;

– сфера застосування розроблюваної НМ відноситься до сфери аналізу текстової інформації;

– пристосованість до автономного функціонування за рахунок автоматизації навчання необов'язкова.

Аналіз наведених умов вказує на їх явно різну значимість при розрахунку оптимального типу нейромережевої моделі. Тому шляхом експертного оцінювання встановлено, що найбільш значимими є умови:

– забезпечення короткого терміну навчання;

– можливості інтерпретації виходу нейронної мережі у графічному вигляді;

– апробованість в задачах аналізу текстової інформації.

Використавши результати [4, 5] визначено, що оптимальним типом неромережевої моделі являється карта Кохонена.

Розробка карти Кохонена

Відповідно [3] розробка моделі карти Кохонена пристосованої для вирішення конкретної задачі складається з таких етапів:

1. визначення номенклатури та допустимих величин вхідних параметрів;
2. підготовка навчальної вибірки;
3. нормування вхідних параметрів навчальної вибірки;
4. визначення кількості кластерів;
5. визначення виду сітки зв'язків між нейронами шару розпізнавання;
6. вибір параметрів навчання;
7. навчання;
8. візуалізація та верифікація результатів навчання;
9. якщо результати не задовільні, необхідно провести навчання з новими параметрами НМ. Для цього повторити п. 4-8.

Таким чином, необхідно сформувати множину вхідних параметрів (п. 1-3), розробити архітектуру НМ (п. 4-6) та провести її навчання (п. 7-9).

Основою формування вхідних даних послужила методика [1, 2, 4], яка передбачає:

1. Формування із піддослідних текстів словника інформативних слів. В словник не включаються малозначущі, службові слова та слова-зв'язки. Наприклад, в словник не включаться слова: в, на, до, а, біля, та.

2. Якщо в сформованому словнику присутні декілька слів-синонімів, то вони замінюються одним словом. Наприклад, слова-синоніми машина, автомобіль, легковик, джип можливо замінити словом автомобіль. Заміна синонімів потребує використання спеціальних граматичних словників, в яких враховується тематика тексту.

3. Розрахунок для кожного із текстів частоти зустрічі кожного із інформативних слів в канонічній формі. Розрахунок частоти зустрічі реалізовано так

$$\mu_i^j = n_i^j / N_i , \quad (5)$$

де μ_i^j – частота зустрічі канонічної форми, n_i^j – кількість всіх словоформ j-го слова в і-му тексті, N_i – кількість слів в і-му тексті.

4. Крім того, до складу вхідних параметрів ввійшли: назва тематики тексту та відносна кількість інформативних слів. Останній параметр розраховується так

$$I_w = I/S , \quad (6)$$

де I_w – відносна кількість інформативних слів реферату листа, I – загальна кількість інформативних слів реферату листа, S – загальна кількість слів реферату листа.

Застосовувався цей параметр для покращення розпізнавання беззмістовних текстів.

Таким чином кількість вхідних параметрів дорівнює

$$N_1 = K + 2 , \quad (7)$$

де N_1 – кількість входів карти Кохонена; K – кількість інформативних слів у навчальній вибірці.

Для розрахунку кількості слів в канонічній формі на мові VBA створено дві програми "Заміна_Словник" та "Заміна". Механізм пошуку словоформ здійснюється з використанням вбудованих в Microsoft Office словників та синтаксичного аналізатора. В програмі "Заміна_Словник" передбачено чотири етапи розрахунків:

1. За допомогою об'єкту Words, бібліотеки Microsoft Word 11.0 Object Library розраховується загальна кількість слів в тексті.

2. Видалення неінформативних слів, що входять до словника, службових слів. Цей словник не входить до складу Microsoft Office, проте його заповнення не викликає труднощів і було здійснене за допомогою [80].

3. Пошук кожного із слів у всіх словоформах. Пошук здійснюється за допомогою об'єкту Words та завершується підрахунком кількості знайдених словоформ і їх видаленням.

4. Запис кожної унікальної словоформи в книгу Microsoft Excel для подальшого аналізу.

Програма "Заміна_Словник" була апробована на декількох текстових фрагментах. В деяких випадках результати були незадовільними через неякісне визначення словоформ. Наприклад, зміна букв в корені слова призводила до його класифікації як нової словоформи. Виправити цей недолік можливо завдяки вдосконаленню синтаксичного аналізатора та словників, що входять до складу Microsoft Office, або завдяки визначенню канонічної форми слова користувачем програми. Через те, що перший варіант виправлення недоліку є досить трудомістким, а програма призначена в основному для експериментальних цілей, було обрано другий варіант. Для цього програма "Заміна_Словник" була дещо модифікована.

Нова програма дістала назву "Заміна". В ній визначення окремої канонічної словоформи здійснюється користувачем шляхом виділення певної групи символів. Після цього програма вже в автоматичному режимі підраховує кількість входжень цієї групи символів. Таким чином, якість визначення канонічних словоформ інформативних слів залежить від кваліфікації та ретельності користувача програми. Програма "Заміна" була апробована та довела свою ефективність при дослідженні текстів рефератів.

Розпізнавання спаму

В якості статистичного матеріалу було використано 100 листів по темі запрошення на семінари, реклами побутових послуг та промислових товарів. Листи були отримані автором на протязі декількох тижнів 2012 року. Можна вважати, що листи однієї із вказаних тематик є цільовими, а інші листи – спам. Попередній аналіз статистичного матеріалу виявив, що кількість слів в канонічній формі в отриманих електронних листах перевищує 1000. В багатьох випадках листи однієї тематики не значно відрізнялись між собою. Наприклад, було отримано 12 листів з запрошенням відвідати семінар по темі "Логістика". Різниця між листами полягала тільки в даті проведення семінару, а перелік інформативних слів залишився незмінним. З точки зору розпізнавання спаму, означені листи повинні відноситись до одного класу. Тому листи з однаковим набором інформативних слів були виділені в окремі групи. Темі "реклама послуг" відповідає група листів № 1, темі "запрошення на семінари" відповідають групи листів № 2, 3, 4, 5, 6, 7, 9, 12, 14, темі "реклама промислових товарів" – № 8, 10, 11, 13. Зміст та умовна класифікація груп листів представлені в табл. 1.

Після розрахунку кількості канонічних форм інформативних слів за допомогою (5) в середовищі Microsoft Excel були розраховані частоти цих слів для кожної із груп отриманих листів. Фрагмент вхідних даних показаний в табл. 2. Зазначимо, що для наглядності в заголовку табл. 2 наведено повну форму інформативних слів.

Таблиця 1

Тематика груп листів

№ групи	Зміст групи листів	Тематика групи листів
1	Реклама супутникової антени	Реклама послуг
2	Семінар по темі "Передача житлового будинку в експлуатацію"	Запрошення на семінари
3	Семінар по темі "Як знайти клієнта телефону?"	
4	Семінар по темі "Кодекс адміністративного судочинства України"	
5	Семінар по темі "Сучасні підходи логістики"	
6	Семінар по темі "Сучасний маркетинг"	
7	Семінар по темі "Сучасний менеджмент"	
9	Семінар по темі "Психологія споживача"	
12	Семінар по темі "Судові спори з податковими органами"	
14	Семінар по темі "Збільшення власного грошового потоку"	
13	Реклама плазмового телевізора	Реклама промислових товарів
8	Реклама обігрівача	
10	Реклама охоронної сигналізації	
11	Реклама систем відеонагляду	

Таблиця 2

Величини вхідних параметрів карти Кохонена

№ групи листів	Відносна кількість інформативних слів	ВАРТІСТЬ	ГРН	ДІМ	ДЕНЬ	РІК
1	0,12	0,041	0,041	0	0	0
2	0,35	0	0,0148	0,051	0	0
3	0,2759	0	0	0	0,053	0
4	0,219	0	0	0	0	0,0238
5	0,2336	0,0263	0,0263	0	0	0,0263

Кластеризація текстів здійснювалось за допомогою пакету Deductor Studio 4.3, на основі загальної методики проектування нейронних мереж, з врахуванням результатів [4]. При побудові карти Кохонена прийнято: розмір топографічної сітки – (16×12), форма сітки зв'язків – гексагон, кількість кластерів – 5, кількість навчальних епох – 500, η=0.1, r=6 на початку навчання, η=0.005, r=1 в кінці навчання. Розділена на кластери карта Кохонена представлена на рис. 1. На рис. 1 межі кластерів показані неперервною лінією, а межі комірок показані пунктиром. Кластери пронумеровані буквами A, B, C, D, E, а групи листів з однаковим набором інформативних слів – цифрами від 1 до 14. Відповідно, цифри 8, 10, 11, 13 відповідають групам листів по темі "реклама промислових товарів", а номер 1 відповідає листам по темі "реклама побутових послуг". Всі інші листи є запрошеннями на семінари. Таким чином, карта Кохонена якісно розділила листи на дві основні теми – реклама (кластери C та D) та запрошення на конференції (кластери A, B, E). Проте не достовірно віднесла до одного кластеру листи з рекламою промислових товарів та листи з рекламою побутових послуг. Однак якість відображення однотипних листів за допомогою карти Кохонена дозволяє провести їх приблизну класифікацію самим користувачем.

Розпізнавання витоків тестової інформації

В якості статистичного матеріалу було використано 30 текстових документів, які входять до бази даних DLP-системи "Контур інформаційної безпеки". Тематика документів: резюме, бухгалтерські звіти та опис технічних характеристик різноманітної побутової техніки. Попередній аналіз статистичного матеріалу виявив, що кількість слів в канонічній формі в цих документах не перевищує 500. Після розрахунку кількості канонічних форм інформативних слів за допомогою (5) в середовищі Microsoft Excel для кожного із документів були розраховані частоти цих слів. Фрагмент вхідних даних показаний в табл. 3.

Як і в задачі розпізнавання спаму моделювання карти Кохонена здійснювалось за допомогою пакету Deductor Studio 4.3 на основі загальної методології проектування нейронних мереж, з врахуванням результатів [4]. При побудові карти Кохонена прийнято: розмір топографічної сітки – (16×12), форма сітки зв'язків – гексагон, кіль-

кість кластерів – 3, кількість навчальних епох – 200, η=0.1, r=6 на початку навчання, η=0.005, r=1 в кінці навчання. Розділена на кластери карта Кохонена представлена на рис. 2.

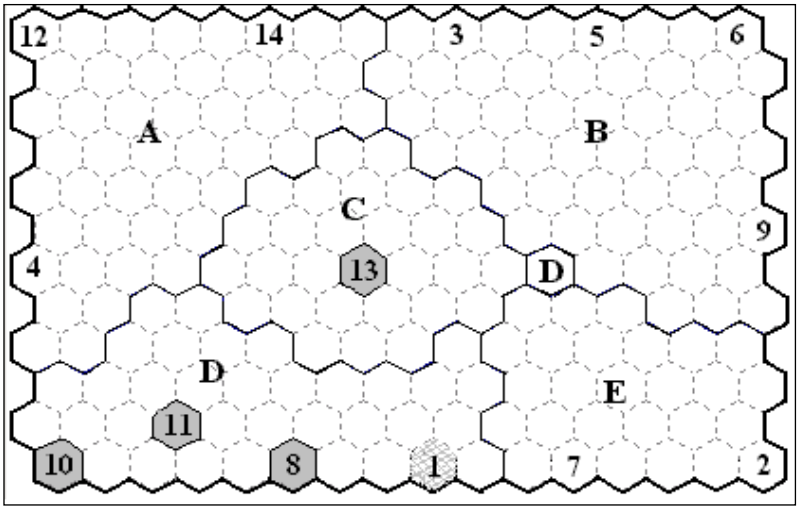

Рис. 1. Карта Кохонена в задачі розпізнавання спаму

Таблиця 3

Величини вхідних параметрів карт Кохонена
та ПК при розпізнаванні витоків інформації

№ листа	Відносна кількість інформативних слів	СТАЖ	ОСВІТА	НАРОДИВСЯ	КОЛІР	ПОТУЖНІСТЬ
1	0,333	0	0,02	0,02	0	0
2	0,212	0,031	0,04	0	0	0
3	0,227	0	0	0	0,01	0,02
4	0,393	0,03	0,014	0,014	0	0
5	0,221	0,04	0,026	0,026	0	0

На рис. 2 межі кластерів показані неперервною лінією, а межі комірок показані пунктиром. Кластери пронумеровані буквами від A, B, C а документи – цифрами від 1 до 30. Кластер A та документи з номерами від 1 до 10 відповідають темі – "резюме", кластер B та документи з номерами від 11 до 20 відповідають "бухгалтерським звітам", а кластер C та документи з номерами від 21 до 30 – "опис побутової техніки". Аналіз рис. 2 дозволяє стверджувати, що карта Кохонена в цілому правильно розділила документи на 3 основних теми: резюме, бухгалтерські звіти та опис побутової техніки. Однак помітні окремі помилки кластеризації. Наприклад, документ № 23, в кому наведено опис телевізора не правильно віднесено до кластеру A який відповідає темі "резюме". Разом з тим, як і у випадку розпізнавання спаму якість відображення однотипних документів за допомогою карти Кохонена дозволяє провести їх приблизну класифікацію самим користувачем.

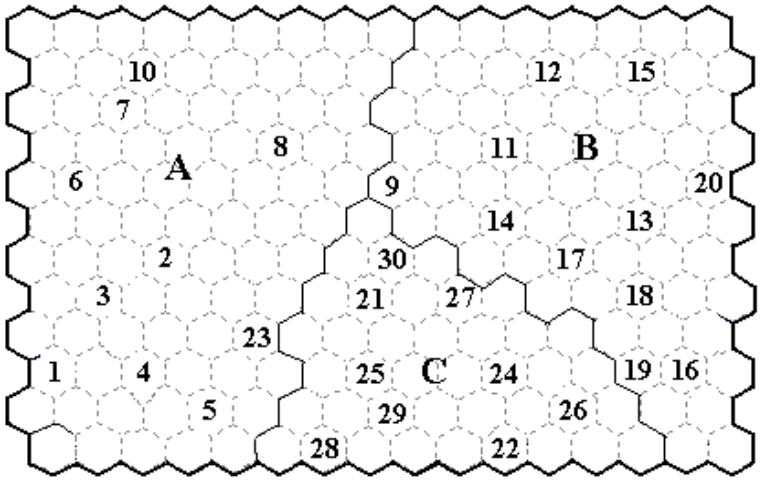

Рис. 2. Карта Кохонена в задачі розпізнавання витоків інформації

Проведені експерименти з метою дослідження впливу параметрів настройки карти Кохонена на відображення топографічного шару та на результати кластеризації. Результати експериментів показали, що відображення топографічного шару на площину головних компонент практично не змінюється. Однак зі зміною параметрів настройки результати кластеризації значно змінюються. При цьому можливо визначити величини параметрів, при яких якість кластеризації залишається задовільною.

Висновки

Вперше запропоновано та обґрунтовано методологію класифікації повідомлень електронної пошти, яка за рахунок використання нейронної мережі типу карти Кохонена надає можливість автоматичного та автоматизованого виявлення спаму та витоків текстової інформації на основі змістовного аналізу тексту.

ЛІТЕРАТУРА

[1]. Заболеева-Зотова А.В. Естественный язык в автоматизированных системах. Семантический анализ текстов / А.В. Заболеева-Зотова. – Волгоград : ВолгГТУ, 2002. – 228 с.

[2]. Зиновьев А.Ю. Визуализация многомерных даннях / А. Ю. Зиновьев. – М. : СК Пресс, 2005. – 180 с.

[3]. Ежов А. А. Нейрокомпьютинг и его применения в экономике и бизнесе / А. А. Ежов, С. А. Шумский. – М. : МИФИ, 1998. – 224 с.

[4]. Терейковський И. А. Применение семантического анализа содержимого электронных писем в системах распознавания спама / И. А. Терейковский // Захист інформації. – 2006. – № 4. – С. 49-60.

[5]. Терейковський І.А. Оптимізація архітектури нейронної мережі призначеної для діагностики стану комп'ютерної мереж / І. А. Терейковський // Наук.-техн. зб. Управління розвитком складних систем Київського національного університету будівництва і архітектури. – 2011. – Випуск 6. – С. 155-158.

REFERENCES

[1]. Zaboleeva-A.V. Zotova Natural language in automated systems. Semantic text analysis / A. Zaboleeva-Zotov., Volgograd Volgograd State Technical University, 2002, 228 p.

[2]. Zinoviev A. Visualization of multidimensional dannyah, M. SK Press, 2005, 180 p.

[3]. Yezhov A. A. Neurocomputing and its applications in economics and business / A.A. Ezhov, SA Shumsky.,Moscow Engineering Physics Institute, 1998, 224 p.

[4]. Tereykovskiy I. A. application of semantic analysis of the content of emails in spam detection systems Zahist informatsii, 2006, № 4, P. 49-60.

[5]. Tereykovskiy I. A. Optimization of neural network architecture designed to diagnose the state of the computer networking, Scientific and Technical. Collected. Management of complex systems Kyiv National University of Construction and Architecture, 2011, Release 6, P. 155-158.

МЕТОДОЛОГИЯ КЛАССИФИКАЦИИ ПИСЕМ ЭЛЕКТРОННОЙ ПОЧТЫ С ИСПОЛЬЗОВАНИЕМ НЕЙРОННЫХ СЕТЕЙ

Эффективность защиты электронной почты во многом зависит от достоверности распознавания в письмах спама и утечек. Существующие средства распознавания базируются на статистических методах анализа текстовой информации, что значительно ограничивает их возможности по выявлению новых видов спама и утечек. С целью преодоления этого недостатка предложена методология классификации писем на основе содержательного анализа электронного с помощью нейронных сетей. Предложено использовать в качестве входных параметров нейронной сети частоту встречи в тексте письма информативных слов в канонические форме. Доказано, что оптимальным типом нейросетевой модели есть карта Кохонена, основным преимуществом которой является высокая скорость обучения и возможность удобной визуализации результатов классификации. Это позволяет быстро реагировать на новые вить спама и утечек и проводить окончательную классификацию писем самим пользователем. Проведенные эксперименты подтвердили эффективность предложенной методологи.

Ключевые слова: электронная почта, спам, утечка, нейронная сеть, карта Кохонена.

CLASSIFICATION METHODOLOGY E-MAILS USING NEURAL NETWORKS

The effectiveness of email security is largely dependent on the accuracy of spam detection in leaves and roots. Existing recognition methods based on statistical analysis of text information, which significantly limits their ability to detect new types of spam and leaks. To overcome this shortcoming proposed classification methodology sheets based on semantic analysis using an electronic neural networks. A used as input parameters of the neural network in the frequency of meeting the letter of informative words in canonical form. It is shown that the best type of neural network model is Kohonen map, the main advantage of which is a high-speed training and the possibility of easy visualization of classification. This allows you to quickly react to new spam and howl leaks and conduct a final classification of letters by the user. The experiments confirmed the possibility of increasing the reliability of detection of 20-30%.

Index Terms: email, spam, leakage, neural networks, Kohonen map.

Терейковський Ігор Анатолійович, кандидат технічних наук, доцент, доцент кафедри системного програмування та спеціалізованих комп'ютерних систем Національного технічного університету України «Київський політехнічний інститут».
E-mail: terejkowski@ukr.net
Терейковский Игорь Анатольевич, кандидат технических наук, доцент, доцент кафедры системного программирования и специализированых компьютерных систем Национального технического университета Украины «Киевский политехнический институт».
Tereykovskiy Igor, Ph.D., associate professor, assistant professor of specialized systems programming and computer systems of the National Technical University of Ukraine "Kiev Polytechnic Institute".

УДК 004.056.5

СТЕГАНОАНАЛИТИЧЕСКИЙ АЛГОРИТМ ДЛЯ ИЗОБРАЖЕНИЙ, ПОДВЕРГАВШИХСЯ ОПЕРАЦИИ СЖАТИЯ С ПОТЕРЯМИ

Владимир Рудницкий, Илья Узун

Легкость в применении, а также масса программных средств как платных, так и бесплатных, свободно распространяемых по сети, сделали стеганографию очень популярным инструментом, за счет которого можно обеспечить простой способ организации утечки ценной информации и неконтролируемый обмен информацией в противозаконных целях. Данные обстоятельства вынуждают активизировать усилия по разработке алгоритмов стеганоанализа. Одним из видов таких алгоритмов являются алгоритмы, базирующиеся на анализе пар цветов цифрового изображения. Большинство существующих подобных средств ориентировано на работу с изображениями, хранимыми в форматах без потерь, что значительно сужает область их применения. С учетом этого, в работе был предложен стеганоаналитический алгоритм определения наличия секретного сообщения, погруженного в цифровое изображение, хранимое в формате с потерями (JPEG), при помощи метода модификации наименьшего значащего бита. Полученные результаты показали, что принцип анализа пар цветов может быть успешно применен и для стеганоанализа цифровых изображений подвергавшихся операции сжатия с потерями.

***Ключевые слова**: стеганография, стеганоанализ, близкие пары цветов, уникальные цвета, сокрытие информации.*

Введение. Стремительное развитие информационных технологий и динамический рост форматов цифровых данных обеспечивают практически неограниченные возможности для сокрытия информации. Одной из наук, занимающихся скрытием данных, является стеганография [4, 5]. Отличительной особенностью стеганографии является то, что при передаче секретной информации в тайне здесь остается сам факт передачи. Преимущество стеганографии состоит в том, что она предоставляет возможность скрытно передать конфиденциальное сообщение – дополнительную информацию (ДИ) одновременно с открытой информацией – контейнером или основным сообщением (ОС), которое не является конфиденциальным. В качестве ОС может быть выбран любой мультимедиа объект – цифровое изображение (ЦИ), видео или аудио (в настоящей работе как контейнер используется ЦИ). В результате погружения ДИ в ОС не должно происходить заметных изменений контейнера. Данный процесс будем называть стеганопреобразованием (СП), а его результат – стеганосообщением (СС). Использование СП часто позволяет избежать прямых атак на ДИ, поскольку неизвестно, присутствует ли она в информационном потоке. ДИ, вносимая в контейнер, может быть предварительно зашифрована, чтобы усложнить задачу стегоаналитика [4]. Основная задача стеганоанализа (СА) [4, 5] – установление факта присутствия в контейнере скрытой информации.

Легкость в применении, а также масса программных средств как платных, так и бесплатных (Steganos, StegHide, S-tools и др.), свободно распространяемых по сети, сделали стеганографию очень популярным инструментом. Это простой способ для организации утечки ценной информации из компаний, неконтролируемого обмена информацией между подозреваемыми и правонарушителями и т.д. Посредством стеганографии между собой общаются как секретные государственные службы, шпионы [15], так и криминальные структуры, и террористы [1, 10, 12]. Поэтому развитие методов СА на сегодняшний день является задачей, актуальность которой трудно переоценить. Работа СА заключается в поиске и анализе определенных характеристик и признаков в исследуемом цифровом объекте, определение