УДК 004.738.5

О. М. Певзнер
Дніпропетровський національний університет ім. Олеся Гончара

АВТОМАТИЧНА КЛАСИФІКАЦІЯ ПОШТОВИХ ЛИСТІВ НА ПІДСТАВІ АНАЛІЗУ ПОВТОРЕНЬ ТЕКСТОВИХ ШАБЛОНІВ

У статті досліджено деякі ефективні алгоритми аналізу електронної пошти щодо наявності в ній спаму. Розглянуто алгоритми Teiresias та Chung-Kwei, виявлено їх переваги та недоліки.

Ключові слова: електронна пошта, спам, алгоритм, шаблон, алгоритм Teiresias, алгоритм Chung-Kwei.

Актуальність проблеми. Нині роль електронних засобів передачі даних в електронному бізнесі дедалі підвищується. У цих умовах електронна пошта посідає особливе місце як один з найважливіших засобів сучасних комунікацій. На жаль, в останні роки у світі простежується негативна тенденція щодо застосування ресурсів електронної пошти в спам-бізнесі. Масштаби спам-бізнесу набувають таких розмірів, які загрожують безпеці Інтернету в цілому, тому завдання боротьби з цим негативним явищем сьогодні, як ніколи, актуальне.

Аналіз останніх наукових досліджень. Існує багато відомих алгоритмів такого розпізнавання [1, 2], та кожний з них має свої переваги та недоліки.

Мета дослідження. Метою роботи є визначення методів об'єктивного розпізнавання спама серед загального поштового трафіка.

Основні результати дослідження. Розглянемо кілька оригінальних алгоритмів, які дозволяють аналізувати поштові відправлення та розрізняти листи за принципом «спам»/«не спам» на підставі аналізу повторень шаблонів у тексті листа.

АЛГОРИТМ TEIRESIAS.

Алгоритм призначений для пошуку в масиві рядків повторень послідовностей (шаблонів).

Нехай є заданим масив рядків в алфавіті A. Будемо вважати за шаблон (у термінах алгоритму Teiresias) послідовність вигляду: $A(A|.)*A$. У цьому разі шаблоном є рядок, який починається та закінчується символами алфавіту A, між якими знаходиться будь-яка комбінація символів алфавіту A та спеціального символу '.' (точка). Шаблон є регулярним виразом, де символ '.' відповідає будь-якому символу алфавіту A. Можна вважати, що відповідний шаблон P визначає мову $G(P)$. Наприклад, якщо заданий шаблон $BC.D.E$, то його мова буде містити, зокрема, такі рядки: $BCCDCE$, $BCEDCE$, $BCBDBEB$ тощо.

Для шаблона P кожний його підрядок, який також є шаблоном, є внутрішнім шаблоном шаблона P. Шаблон P будемо називати (L,W)-шаблоном, якщо кожний його внутрішній шаблон завдовжки W або більше містить як мінімум L символів алфавіту A. Зрозуміло, якщо шаблон P є (L,W)-шаблоном, то він також є й $(L,W+1)$-шаблоном.

Рядок символів алфавіту A підпадає під шаблон P, якщо він містить підрядок з мови $G(P)$. Якщо заданий набір рядків $S=\{s_1,s_2,...,s_n\}$, то для шаблона P можна визначити таку множину зміщень:

$L_s(P)=\{(i,j)|$ рядок s_i містить рядок мови $G(P)$, починаючи зі зміщення $j\}$.

Шаблон P' є більш характерним за P, якщо він може бути отриманий з P шляхом заміщення одного або кількох спеціальних символів '.' на символи алфавіту A або через дописування праворуч або/та ліворуч рядків, які складаються зі спеціальних символів та символів алфавіту A. Очевидно, що $|L_s(P')| \leq |L_s(P)|$.

Шаблон P будемо називати максимальним для множини S, якщо не існує більш характерного шаблона P', такого, що $|L_s(P')|=|L_s(P)|$.

Алгоритм Teiresias дозволяє за множиною рядків $S=\{s_1,s_2,...,s_n\}$ в алфавіті A та параметрам L, W, K відшукати всі максимальні (L,W)-шаблони, під які підпадають як мінімум K різних рядків множини S. Детальний опис цього алгоритму можна знайти в [3, 4]. Спробувати алгоритм у роботі можна на сайті [5].

АЛГОРИТМ CHUNG-KWEI.

Алгоритм Chung-Kwei базується на застосуванні алгоритму Teiresias для пошуку шаблонів в електронних повідомленнях. Цей алгоритм є цілком евристичним.

Листи розглядаються як набір рядків в алфавіті ASCII. Автори передбачають розподіл листів на дві частини: технічна інформація (заголовки) та тіло листа. Пропонується застосовувати відповідний алгоритм до кожної частини листа окремо. Алгоритм виконується у два етапи: створення бази шаблонів (навчання) та застосування бази шаблонів для класифікації листа.

Для створення бази шаблонів (словника спама) використовується початковий набір спама, до якого застосовується алгоритм Teiresias з певними значеннями (L,W) і K=2. Зрозуміло, що отримані шаблони є об'єктивними характеристиками документів та можуть розглядатися як база шаблонів для будь-якого іншого методу автоматичної класифікації, наприклад, в класифікаторах Байєса [1]. Якщо крім набору спама в наявності є також набір легальної пошти, то шаблони виділяються й з нього. Ці шаблони можна використовувати у подальшому для того, щоб видалити зі словника спама зайві шаблони і таким чином зменшити ймовірність помилкових спрацьовувань.

Після отримання словника спама можна виконувати класифікацію листів. Вона полягає в тому, що в листі шукають надходження шаблонів зі словника спама. Якщо кількість знайдених шаблонів невелика (менша за наперед заданий поріг), то класифікація припиняється, і лист вважається за легальний.

Для кожного символу листа, що обробляється, встановлюється окремий лічильник, який на початку обнуляється.

Кожне надходження шаблона до листа також відповідає надходженням цього ж шаблона в певні листи початкового набору. Для кожного такого надходження за таблицею відповідностей символів нараховуються очки до лічильників. Наприклад, лист містить підрядок $ABCD$, який відповідає знайденному шаблону, та база шаблонів містить підрядок $AbCD$. Тоді кожному з чотирьох лічильників, які відповідають символам цього рядку, додаються очки для пар символів (A,A), (B,b), (C,C) та (D,D). Таблиця відповідностей символів заповнюється на початку обробки, виходячи з прагматичних міркувань про ступінь «схожісті» символів. Якщо після завершення обробки шаблонів відсоткова кількість ненульових лічильників (покриття листа шаблонами) виявиться невеликою (менша за наперед задане порогове значення), то лист вважатиметься за легальний. В іншому випадку лист буде класифікований як спам.

Знайдені спамерські листи можуть бути автоматично додані до відповідної бази та використовуватимуться в подальшому як елемент навчання алгоритму.

Статистика застосування метода Chung-Kwei на великих потоках електронної пошти показує, що результати класифікації спама становлять близько 96% правильного розпізнавання спама при 0,6% помилок [6]. Це – один з кращих результатів розпізнавання спама на сьогодні. Більш детальний опис алгоритма Chung-Kwei та результатів його тестування можна знайти в [6].

Висновки та перспективи подальших досліджень. Значна кількість відомих сьогодні алгоритмів класифікації спама насамперед виявляється орієнтованою на кінцевого користувача – безпосереднього отримувача електронної кореспонденції. Проте слід розуміти, що найбільша частка спама припадає на поштові сервери. Саме вони виявляються найбільш незахищеними від спама та зазнають найбільших збитків від нього. У той самий час алгоритми захисту від спама, які можуть використовуватися кінцевими користувачами, принципово відрізняються від тих, що мають серверне застосування. Розглянутий нами алгоритм Chung-Kwei, який базується на методі Teiresias, є саме таким, що може бути рекомендований як серверний. Ефективність цих алгоритмів прямо залежить від розмірів поштової бази, яка використовується під час навчання; ця база має бути досить великою та різноманітною. Крім того, передбачається, що всі листи, на яких здійснювалося навчання, мають залишатися доступними під час класифікації поштових повідомлень (за ними розраховується покриття шаблонами відповідного листа). Якщо прийняти середній розмір листа в 5Кб, то база з 65 тис. листів вимагатиме більш ніж 300Мб на жорсткому диску. Ясно, що жодний користувач не стане зберігати у себе таку величезну базу спама та постійно оновлювати її.

Зазначимо також, що метод Chung-Kwei, на відміну від відомого методу Байєса та деяких інших, не має проблеми «перенавчання»: ефективність його застосування прямо залежить від розмірів та різноманітності навчальної бази. У той самий час слід розуміти, що саме поняття спама у серверному розумінні є невизначеним, тому створення баз спама та легальної пошти являє собою досить складну та нетривіальну задачу. За цих умов алгоритм Chung-Kwei не можна вважати універсальним, але його застосування на поштових серверах сьогодні слід визнати найкращим вибором внаслідок його високої ефективності.

Бібліографічні посилання і примітки

1. Певзнер О. М. Метод байєсівської фільтрації спаму / О.М. Певзнер, В.В. Борщевський. – Д.: ДНУ, 2004. – С. 23–28. (Матеріали І Всеукраїнської студентської наукової конференції 19–20 березня 2004 р.).
2. Певзнер О. М. Моделювання та аналіз ефективності зниження спам-ризиків за допомогою марківської фільтрації / О. М. Певзнер. – Кривий Ріг, 26–28 квітня 2005 р. – С. 156–164. (Матеріали VI Всеукраїнської науково-практичної конференції «Комп'ютерне моделювання та інформаційні технології в науці, економіці та освіті»).
3. Rigoutsos I. Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm / I. Rigoutsos, A. Floratos // Bioinformatics, 1998. – Vol. – 14. – no. 1.
4. Floratos A. Research report: On the time complexity of the TEIRESIAS algorithm / A. Floratos, I. Rigoutsos // IBM Research division, RC21161(94582)21APR98, 1998.
5. IBM Bioinformatics Group – Tools & Content. – Режим доступу: http://cbcsrv.watson.ibm.com/Tspd.html
6. Rigoutsos I. Chung-Kwei: a pattern-discovery-based system for the automatic identification of unsolicted e-mail messages (SPAM) / I. Rigoutsos, T. Huynh

Надійшла до редколегії 30.06.2009.