УДК 004.93`11

Галла Ю.В.

АНАЛІЗ ЗАСТОСУВАННЯ МЕТОДІВ РОЗПІЗНАВАННЯ ОБРАЗІВ ДЛЯ ВИЯВЛЕННЯ ЕЛЕКТРОННИХ РЕКЛАМНИХ РОЗСИЛОК

Інститут комп'ютерних технологі

В статті розглянута проблема застосування методів розпізнавання образів при автоматичній обробці електронних повідомлень для виявлення серед них небажаної кореспонденції рекламного характеру, що може бути класифікована як „спам". Ця проблема є специфічною і безпосередньо пов'язана з розвитком та вдосконаленням електронної пошти, як зручного засобу комунікації, що активно використовується для просування товарів та послуг. У статті представлений огляд існуючих масштабів проблеми небажаних розсилок; наведена класифікація методів, що застосовують „спамери" для розсилок; здійснено огляд методів обходу автоматизованих фільтруючих систем та методів і алгоритмів існуючих „антиспам-систем"; запропоновано використання адаптивного методу розпізнавання образів для обробки „графічного спаму".

Вступ

Електронна пошта являє собою сучасний і високотехнологічний засіб комунікації. На теперішньому етапі розвитку інформаційних технологій, цей комунікаційний канал активно використовується не тільки для обміну інформацією, але й для просування товарів і послуг, у тому числі й для проведення масових анонімних незапрошених рекламних кампаній (іншими словами - для розсилання „спаму").

Інші види масових розсилок (підписні листи та ін.) не можуть бути так класифіковані, тому що користувач дає на них згоду (запрошені розсилки).

Явище, назване „спамом", існує вже тринадцять років [1]. За час що пройшов були розроблені спеціалізовані технології, випущені програми для небажаних розсилок, відбувається тісна співпраця „спамерів" з комп'ютерними злочинцями та розробниками вірусів для більш ефективного розповсюдження „спаму".

Згідно досліджень, що наведені у [1], ще у 2004 році були виявлені такі тенденції у розвитку „спаму":

– криміналізація „спамерского" бізнесу і посилення законодавчої боротьби зі „спамом";

– становлення тематичних лідерів „спаму";

– ріст збитків, що на кінець 2004 року оцінювався як мінімум 250 млн. евро тільки для провайдерів та користувачів Росії;

– застосування високотехнологічних прийомів розсилок;

– ріст обсягів спаму до 75% - 85% від загального поштового трафіку й стабілізація обсягів на цьому значенні, за даними [2] (див. рис.1);

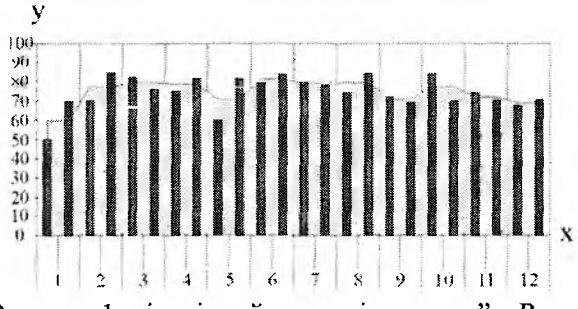

Рисунок 1: кількісний розподіл „спаму" в Рунеті за 2005 р. Вісь *X* – місяці року; вісь *Y* – відсоткова доля „спаму" у пошті

Попередні дослідження

Класифікація „спаму" потрібна для виявлення його основних ознак, що будуть використані системою розпізнавання образів.

Основні признаки „спаму" можна розділити за типом листа та його змістом. За типом листів до ознак „спаму" відносяться:

– розсилки не ініційовані користувачем (анонімні, на відміну від періодичних підписних Інтернет видань, наприклад *Subscribe.ru*);

– листи малого розміру – до 100 Кб;

– розсилки, що здійснюється багатомільйонним тиражем за короткий час (для порівняння: 2004 – близько 24 годин, 2006 – 20-30 хв.);

– розсилки, що не дозволяють однозначно ідентифікувати відправника.

За змістом листа „спам" можна розподілити (за інформацією [2]) на такі категорії, див. рис. 2:

а) комп'ютерне шахрайство; б) реклама порно-сайтів та послуг/товарів „для дорослих"; в) реклама контрафактних ліків; г) реклама освіти; д) комп'ютерні та Інтернет послуги; е) особисті фінанси (страхування, вкладання коштів); ж) реклама відпочинку та туризму; з) реклама послуг електронної реклами; і) інші товари та послуги;

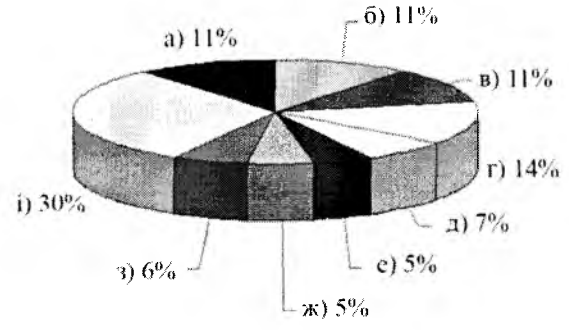

Рисунок 2. Розподіл тематик „спаму" у Рунеті за 2005 рік

Найрозповсюдженіші технологічні рішення „спамерів"

Серед новинок і успіхів „спамерів" можна відзначити:

– чітка схема розсилки – вона складається з декількох етапів, що можуть виконуватися незалежно один від одного. Згідно [2] етапи є такими: 1) збір та верифікація e-mail адрес, класифікація адрес *по типам; 2) підготовка точок розсилки;* 3) створення ПЗ для розсилки; 4) пошук клієнтів; 5) створення рекламних оголошень;

– фішинг – „новинка" 2004 р. Це розповсюдження підробних повідомлень від імені банків/фінансових компаній, ціллю яких є збір паролів, пін-кодів та іншої конфіденційної фінансової інформації;

– фармінг - „новинка" 2005 р. Це підміна *URL* сторінки, у результаті якої користувач опиняється на шахрайському сайті, що повністю копіює вигляд оригінальної сторінки легітимного сайту-мішені і залишає там свою конфіденційну інформацію;

– ботнет – використання для розсилок мереж зомбі-комп'ютерів. Кількість комп'ютерів у такій мережі коливається від десятків до сотень тисяч (рекорд – 1.5 млн комп'ю-терів, заражених троянською програмою віддаленого адміністрування *Backdoor. Win32.Codbot*);

– швидкісні розсилки – розсилка здійснюється за 20-50 хвилин з розрахунком обігнати регулярні поновлення баз даних „антиспамерських" програм;

– розсилка зі зворотнім зв'язком – базується не на швидкості, а на модифікаціях зовнішнього вигляду повідомлень;

– графічний „спам" – повідомлення представляється у вигляді графічного зображення.

Тенденції розвитку „спаму"

2005-2006 роки можна охарактеризувати як стабілізацію „спам"-індустрії.

Його частка у потоці пошти близько 80 % і росте пропорційно загальному об'єму всієї кориспонденції.

„Спам", як бізнес визначився у своїй економічній ніші і впевнено зайняв її. Ця справа стала непосильним завданням для новачків і потребує ретельної технічної підготовки та висококваліфікованого персоналу.

Продовжується криміналізація „спаму" – фішинг та ін.

Технологічний ланцюжок розсилок став фіксованим для більшості „спамерів". При організації розсилок визначився чіткий розподіл праці і кожним з етапів підготовки до атаки займаються підготовані саме до цього групи.

Методи боротьби зі "спамом" та задачі, що стоять перед „антиспам"-системами

З 2004 р. на більшості великих публічних поштових сервісів та у більшості провайдерів були встановлені антиспам-системи. Ці системи дозволяють фільтрувати до 90% „спаму". Таким чином, має місце технологічна боротьба між розробниками спамерського та антиспамерського ПЗ.

Всі фільтри використовують в цілому подібні методи визначення „спаму”. Здійснимо огляд деяких методів.

Контентні фільтри використовують розбір технічних атрибутів листа, пошук характерних термінів, сигнатурний метод (розсилка блокується ще до надсилання всіх листів). Класичним прикладом є *Symantec Brightmail Anti-Spam* чи *Kaspersky Anti-Spam.*

Евристичні методи. Найбільш відомим представником є програма SpamAssassin. Існує біля 2000 правил, котрі тестують лист на приналежність до „спаму”.

Технологія цифрових відбитків. Для ідентифікації „спаму” призначається контрольна сума для тексту або графічного зображення листа. Ця контрольна сума заноситься у базу даних і дуже часто поновлюється.

Технологія виявлення спуфінга – виявляються листи з фальшивим відправником.

„Чорні списки” *IP*-адрес – *DNSBL.* Особливо такі системи розповсюджені у провайдерів у сукупності з контентним фільтром. Представники – *Trend Micro RBL+, Spamhaus.*

Ще один підхід оснований на аналізі масовості повідомлення. У Росії його використовує „Яндекс.Пошта”, а у США зустрічається у продуктах компанії *Commtouch.* [3]

Існують різні прийоми обману автоматизованих систем фільтрації. Зупинимося на контентних методах. Вони, у своїй більшості [4], стосуються принципів формування тексту листів, тому що антиспам-системи при аналізі найбільше уваги приділяють тексту повідомлення. На сьогоднішній день проста розсилка однакових листів є повністю неефективною – вони будуть гарантовано виявлені фільтрами по критерію масовості. Перед зловмисниками стоїть задача створити декілька варіантів одного й того ж рекламного тексту, котрі будуть задовольняти наступні умови:

– програма фільтрації буде сприймати ці варіанти як різні та не пов’язані з рекламою;

– людина, що читає листа сприйме ті ж самі варіанти листа як однакові та рекламу.

Основними технологіями „індивідуалізації” повідомлень, як слідує з [4], є навмисні перекручування тексту.

Внесення випадкових текстів, „шуму”, невидимих текстів. Випадковий текст вноситься в початок чи кінець листа, набирається дуже малим шрифтом або кольором фону. Ці додатки утруднюють роботу нечітких алгоритмів та статистичних методів. У відповідь на це з’явився пошук цитат, детальний аналіз *HTML* та інші методи поглибленого аналізу змісту листа.

Перефразування текстів. Одне й те ж повідомлення розсилається у множині варіантів одного й того ж тексту (наприклад, деякі чи всі фрагменти листа замінюються на доречний синонім). При такому прийомі факт перефразування можна встановити тільки після одержання більшої частини розсилки і тільки тоді ефективно налаштувати фільтр.

Навмисні перекручування на рівні слів. Типологію таких перекручень можна розбити на три типи: заміна одних елементів слова (букв) іншими; розбиття слова небуквеними символами; вставка у слово „зайвих” елементів. Такі маніпуляції мають своєю ціллю перешкодити автоматичному аналізатору ідентифікувати приналежність тексту певній тематиці.

Психолінгвістичні та нейролінгвістичні методи формування повідомлення. Текст листа маскується під особисте повідомлення [5].

Представлення всього листа у вигляді графічного зображення („графічний спам”). Новинкою середини 2006 р. став анімаційний „спам” [6].

Тема ефективної фільтрації графічного „спаму” є мало вивченою у даному контексті і потребує додаткового аналізу. Завданням аналізу є оцінка можливості застосування методів розпізнавання образів для автоматичної фільтрації графічного „спаму”.

Опис загальних алгоритмів фільтрації

Графічні листи з'явилися у 2004 р. і займали близько 15% від всього „спаму". За підсумками 2006 р. відмічається великий ріст цього виду листів. Це пов'язано з тим, що тільки деякі фільтри готові аналізувати графічні зображення.

Розсилка однакових графічних листів, як і звичайних „спам"-листів, є повністю безперспективною, тому що фільтри розпізнають таку розсилку за ознакою масовості. Тому „спамери" вносять у графічну картинку „шум" – застосовують змінний фон, фонові малюнки, хвилястий текст, комбінації різних шрифтів, навмисні перекручування тексту, ділять весь малюнок на випадкові частини. У 2004 р. поява варіативного (зашумленого) „спаму" на 2 місяці стала дуже великою проблемою для антиспам-систем. У 2006 р. спамери знову зробили ставку на графіку [6]. Анімаційний „спам" виявився найуспішнішим їх технологічним рішенням.

Спамери використовують *GIF*-анімацію, тому що вона автоматично розпізнається і відтворюється всіма популярними браузерами. Перші розсилки містили від 2 до 4 кадрів, серед яких тільки один кадр був значимим – саме у ньому відтворювався текст. Інші кадри містили фон та решту елементів малюнка, що не несуть змістового навантаження. У подальших розсилках збільшувалась кількість кадрів (більше 20), а також текст реклами розносився по різним кадрам (інформативний не один кадр, а 1-2 десятки).

Застосування алгоритмів розпізнавання образів.

Графічний формат листа дуже ускладнює завдання фільтрів та збільшує вірогідність помилки – кваліфікування легального листа як „спаму". Тому задачу розпізнавання потрібно розбити на декілька етапів і застосувати комплекс алгоритмів, що аналізують різні частини листа на приналежність до групи небажаних розсилок.

На першому етапі аналізу повідомлення, потрібно кваліфікувати лист як підозрілий за будьяким з алгоритмів. Наприклад, у системі *Kaspers-ky Anti-Spam* про кожне повідомлення складається блок інформації, до якого входить інформація: розмір листа; шлях, по якому лист потрапив до адресата (*IP*-адреси останніх поштових релеїв); *md5*-хеші, прораховані для декількох варіантів перетворення вихідного тексту листа (видалення стопслів, нормалізація орфографії, сортування слів по кількості входжень і т.п.); сигнатури обробки графічних включень. Ця інформація не дозволяє відновити зміст листа, однак вона дозволяє групувати однакові поштові повідомлення.

Підозрілі повідомлення, де були помічені графічні фрагменти чи повністю складаються з *GIF*-файлу, повинні аналізуватися окремо.

GIF-файл має блочну структуру, блоки у більшості випадків ніяк не пов'язані між собою, йдуть один за другим та створюють анімацію. Кількість блоків анімації форматом файлу не обмежується і значима інформація повідомлення може бути рознесена по абсолютно всім кадрам, тому потрібно організувати програмний буфер, у якому можна сумішати блоки у потрібній послідовності і аналізувати отримані графічні зображення на предмет присутності тексту. При цьому потрібно враховувати можливу присутність у графічному зображенні навмисних спотворень символів і зашумленості.

На другому етапі аналізу потрібно використовувати методи, котрі виділяють на графічному зображенні місця передбачуваного розташування текстових символів – наприклад використавши хвильовий алгоритм встановлення скелету растрового зображення. При цьому потрібно провести попередню обробку зображення для виявлення характеру шуму та перекручувань і їх усунення. Хвильовий алгоритм полягає у аналізі шляху сферичної хвилі по зображенню та встановлення контурів символів. Потім здійснюють встановлення скелету символів та їх сегментацію (фрагментацію графічного образу рядка тексту на окремі символи). При скелетизації та сегментації можна застосовувати топологічний опис зображень, як пласких графів і цікавитися тільки їх зовнішніми

та внутрішніми контурами. Задача розпізнавання може бути зведена до встановлення гомеоморфності знайденого зображення з одним з еталонних. Її можна виявити за допомогою топологічних інваріантів – таких властивостей зображення, котрі не змінюються при його гомеоморфних перетвореннях. Важливою перевагою топологічного опису є стійкість до сильних деформацій зображення.

На третьому етапі потрібно застосувати адаптивне розпізнавання символів, що виділені на попередньому етапі.

На четвертому етапі вже можуть застосовуватися автоматичні методи ідентифікації змісту листа та механізми фільтрації спаму, що здатні працювати з актуальним поштовим потоком.

Опис методу адаптивного розпізнавання символів

Адаптивний алгоритм [8] базується на синтезі двох підходів до розпізнавання – шрифтового та шрифтонезалежного. Кожен з них має свої переваги та недоліки, а їх поєднання веде до суттєвого підвищення якості розпізнавання.

Математична модель адаптивного розпізнавання. Модель охоплює два ключових етапи адаптивного розпізнавання: кластеризація символів навчаючої виборки та дорозпізнавання. Оцінимо теоретичну межу якості розпізнавання і надійність при заданих параметрах первинного розпізнавання і міри зашумленості символів. Параметри моделі:

P – якість розпізнавання, одержана на етапі первинного розпізнавання;

σ – міра спотвореності символів, дає числовий вираз кількості випадкових змін у конфігурації пікселів серед екземплярів символів, що позначають одну й ту ж букву;

F – фінальна якість розпізнавання за допомогою шрифтозалежного алгоритму, що адаптований до даної вибірки символів;

V – надійність розпізнавання символу;

$V = f(x, P)$, x – відстань від даного символу до центру кластера (ідеального символа). Функція f є частиною конкретного алгоритму обчислення відстані між символом та кластером, тому точність залежить від матеріалу, на якому зібраний кластер.

Припустимо обрана метрика – функція, що відображує відмінності між символом та кластером у дійсне додатне число (відстань). Основним положенням моделі є те, що відстань від символа, що розпізнається до кластера є нормально розподілена випадкова величина з густиною ймовірності $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-x^2}{2\sigma^2}}$

Тоді за заданою мінімально допустимою надійністю V_{min} обчислимо максимальну відстань X_m на яку символ може відхилитися від кластера і при якому $V \geq V_{\min}$

$$X_m = f^{-1}(V_{\min}, P)$$

Далі за визначенням функції розподілу одержуємо [7]: $F = \int_0^{X_m} p(x)dx$. Цей вираз визначає якою буде якість розпізнавання при заданих надійності та ступеню спотворення символів.

Великий інтерес викликає вимірювання величини σ - середньоквадратичного відхилення, тому що вона надає числового виразу поняттю „якість тексту". В цій моделі σ набуває конкретного фізичного змісту – описує варіації котрі виникають у конфігурації пікселів, що описують оригінал символа у процесах друку та сканування. Застосування шкали, що базується на мірах відхилення знаходить своє застосування у різних аспектах розпізнавання, зокрема:

– верифікація результатів кластеризації (кластер з відхиленням, що суттєво відрізняється від середнього по вибірці повинен викликати підозру і бути додатково перевірений);

– динамічне налаштування різних порогових констант, що керують розпізнаванням;

– екстремальні значення σ свідчать про те, що сама адаптація до даної вибірки є невигідною бо необхідна статистична інформація в ній відсутня.

Цікавим із практичної точки зору є питання про те, наскільки близькі параме-

три реального кластера до параметрів розробленої моделі. Нижче приводиться оцінка, що дозволяє визначити як з ростом кількості символів параметри кластера сходяться до теоретичного.

Візьмемо довільний елемент кластера. Нехай p - імовірність появи тут чорного піксела при черговому додаванні символу в цей кластер. Очевидно, що ця ймовірність фіксована самою моделлю й залежить тільки від положення елемента всередині сітки. У такий спосіб процес появи чорних пікселів у даному елементі задовольняє схемі випробувань Бернуллі. У процесі фізичної реалізації потрапляння символів у кластер у цьому елементі існує ξ - частота потрапляння сюди чорного піксела. Це випадкова величина, зосереджена біля p і по центральній граничній теоремі [7] відхиляється від неї відповідно до нормального закону розподілу, отже:

$$-x_\alpha \le \sqrt{\frac{N}{p(1-p)}}(\xi - p) \le x_\alpha,$$

де x_α - квантиль рівня α; α - рівень значущості; N – кількість символів у кластері.

Ця нерівність виконується з ймовірністю $1-2\alpha$. Спростимо нерівність, враховуючи що $p(1-p) \le 1/4$.

$x_\alpha \ge |\xi - p| \times 2\sqrt{N}$ з ймовірністю не меншою за $1-2\alpha$. Припустимо, що кількість символів N=121 (приблизна кількість однакових букв на друкованій сторінці) та припустимо, що $|\xi - p| \le 0.07$, тоді $x_\alpha = 0.07 \times 2 \times 11 = 1.54$ це відповідає рівню значимості 0.0618. Це означає, що наше припущення виконується з імовірністю не меншою за $1-2\times 0.0618 \cong 0.88$. В цих міркуваннях не накладалось ніяких специфічних умов на елемент, отже висновок справедливий для всіх елементів даного кластера. Таким чином можна стверджувати, що при вказаній місткості кластера у майже 90% його елементів абсолютна похибка відхилення від моделі складе не більше 0.07. При розробці конкретної процедури обчислення відстані до кластера або надійності постає питання про коректність порівняння фізично одержуваних значень із константами, що обчислені за допомогою математичної моделі. Маючи подібний механізм, можна виміряти й компенсувати некоректність такого порівняння.

Розглянемо схему роботи адаптивного розпізнавання. Функціонування схеми розділяється на кілька етапів: первинне розпізнавання, збір статистики, кластеризація зібраної статистики, формування еталонів (бази характеристик), до розпізнавання.

Коротко про кожний з названих етапів.

– первинне розпізнавання означає розпізнавання всієї сторінки за допомогою шрифтонезалежного алгоритму;

– збір статистики має на увазі процес відбору надійно розпізнаних символів, які згодом складуть навчальну вибірку для шрифтозалежного алгоритму;

– кластеризація - розбивка навчальної вибірки на кластери(класи). За допомогою такої розбивки уточнюються результати розпізнавання, отримані на етапі первинного розпізнавання, буде виявлена статистична структура сторінки, тобто отримана відповідь на питання: чи групуються однакові символи на даній сторінці, підготовлений вихідний матеріал для навчання шрифтозалежного алгоритму [9];

– верифікація символів-кандидатів у навчальну вибірку за допомогою незалежного методу. Наприклад, словниковий контроль, частотні двобуквені та трибуквені сполучення. Це потрібно для зниження кількості помилок при формуванні еталонів розпізнавання;

– формування еталонів це створення остаточних, двійкових наборів даних (бази характеристик), по яких буде виконуватися дорозпізнавання;

– дорозпізнавання - другий прохід розпізнавання по всій сторінці з метою уточнити результати первинного розпізнавання, виставити адекватні оцінки точності, дорозпізнати те, що було не розпізнане раніше, відзначити ненадійно розпізнані символи [10].

Висновки

У відповідності з викладеним можна

зробити наступні висновки:

1. Методи розпізнавання образів для фільтрації спаму успішно застосовуються багатьма поштовими серверами у світі. Графічний формат листів знижує швидкість аналізу змісту розсилань і тому ефективність фільтрації залежить від повноти комплексу методів, що застосовані на конкретній поштовій системі.

2. Проведено огляд існуючих масштабів проблеми небажаних розсилок; наведена класифікація методів, що застосовують „спамери" для розсилок; здійснено огляд методів обходу автоматизованих фільтруючих систем та методів і алгоритмів існуючих „антиспам-систем"

3. Запропоновано та обґрунтовано застосування адаптивного алгоритму розпізнавання у „антиспам-системах".

4. Розроблена математична модель методу
адаптивного розпізнавання символів, оцінена теоретична межа якості розпізнавання і надійність при заданих параметрах первинного розпізнавання і міри зашумленості символів.

5. Наголошено на невирішеності багатьох питань, за якими потрібно продовжувати дослідження: не повністю вирішеною є проблема центрування символу при накладенні його на сітку еталона при дорозпізнаванні, і взагалі проблема знаходження реперних точок, що точно описують положення растра; суттєвою також є сама проблема дорозпізнавання - взаємодії двох незалежних алгоритмів розпізнавання і вирішення конфліктів між ними; окремого вивчення також потребує проблема розпізнавання символів у які внесені навмисні викривлення та шум.

Список літератури

1. *Ашманов И., Власова А., Тутубалин А.* Спам 2004: аналитический отчет. ЗАО "Ашманов и Партнеры". – М., 15.01.2005. http://www.spamtest.ru

2. *Власова А., Зоркий К., Калинин А.* Спам 2005: аналитический отчет. «Лаборатория Касперского». – М., 10.01.2006. http://www.spamtest.ru

3. *Дронов В.* Весь этот спам. «Лаборатория Касперского». – М., 30.03.2006. http://www.spamtest.ru

4. *Власова А., Зоркий К.* Проблема намеренных искажений письменного текста в электронных рекламных рассылках (спаме). ЗАО «Ашманов и Партнеры». – М., 14.06.2004 http://www.spamtest.ru

5. *Власова А.* Спам основные тенденции в третьем квартале 2006 года. «Лаборатория Касперского». – М., 26.10.2006. http://www.spamtest.ru

6. *Власова А.* Спам в 2006 году: основные вехи. Материалы 4-й международной конференции «Проблема спама и ее решения»

7. *Розанов Ю.А.* Теория вероятностей, случайные процессы и математическая статистика. М. : "Наука", 1989

8. *Ян Д.Е., Анисимович К.В., Шамис А.Л.* Новая технология распознавания символов. Теория, практическая реализация, перспективы. М. : Препринт, 1995

9. Сборник Классификация и кластер. М. : "Мир", 1980

10. *Y-H Pao* Adaptive pattern recognition and neural network "Addison-Wesley" 1989