Посилання на статтю

Димо О.Б. Використання нейронної мережі Кохонена у проектах розпізнавання рекламних текстів / О.Б. Димо, Г.С. Морозова // Управління проектами та розвиток виробництва: Зб.наук.пр. – Луганськ: вид-во СНУ ім. В.Даля, 2010. – № 4(36). – С. 44-50.- Режим доступу: http://www.pmdp.org.ua/images/Journal/36/10dobrrt.pdf

УДК 681.5

О.Б. Димо, Г.С. Морозова

ВИКОРИСТАННЯ НЕЙРОННОЇ МЕРЕЖІ КОХОНЕНА У ПРОЕКТАХ РОЗПІЗНАВАННЯ РЕКЛАМНИХ ТЕКСТІВ

Розглянуто задачу побудови комп'ютерної системи розпізнавання рекламного характеру тексту для фільтрації текстів в Інтернеті. Рис. 6, дж. 3.

Ключові слова: нейронна мережа, Кохонен, розпізнавання, рекламний текст.

А.Б. Дымо, А.С. Морозова

ИСПОЛЬЗОВАНИЕ НЕЙРОННОЙ СЕТИ КОХОНЕНА В ПРОЕКТАХ РАСПОЗНАВАНИЯ РЕКЛАМНЫХ ТЕКСТОВ

Рассмотрена задача построения компьютерной системы распознавания рекламного характера текста для фильтрации текстов в Интернете. Рис. 6, ист. 3.

A.B. Dymo, A.S. Morozova

KOHONEN`S NEURON NETWORK FOR THE ADVERTISING TEXT RECOGNITION PROJECT

Task of construction the computer system of recognition the publicity character of text I order to filtrate texts in the Internet is considered.

Постановка проблеми. Сьогодні в мережі Інтернет є велика кількість небажаної інформації. Задача її ідентифікації і фільтрації вирішується сьогодні багатьма методами, але кожний з них не забезпечує необхідну якість розпізнавання. Більшість з такої інформації є рекламними текстами, тому першочерговою проблемою є їх ідентифікація, тобто розпізнавання. В даній роботі механізмом реалізації системи розпізнавання пропонуються нейронні мережі. Нейронні мережі моделюють роботу людського мозку, що є найбільш ефективним з відомих механізмів розпізнавання.

Метою дослідження є підвищення ефективності, удосконалення, аналіз комп'ютерної системи і механізму реалізації для ухвалення рішення про "рекламності".

Основна частина. Людина, як надзвичайно складна інформаційна система, здатна розпізнавати образи, описи об'єктів. У процесі розпізнавання певної інформації у мозку людини відбувається верифікація інформаційних потоків у сигнали, нервові імпульси, які передаються корі півкуль головного мозку за допомогою активації специфічних клітин, нейронів. Нейрони діляться на збудливі (тобто активуючі розряди інших нейронів) і гальмівні (що перешкоджають

збудженню інших нейронів). Кожен нейрон має довгий відросток, аксон, по якому він передає імпульси іншим нейронам. Аксон розгалужується і в місці контакту з іншими нейронами утворює синапси (місце контакту двох нейронів) на тілі нейронів і дендритів (коротких відростках). Таким чином, один нейрон приймає сигнали від багатьох нейронів і у свою чергу посилає імпульси іншим. Отже нервовий імпульс, залучаючи певні групи нейронів, активує відповідні зони головного мозку людини, які відповідають за ті чи інші моторні реакції, зорове, просторове сприйняття тощо.

Розуміння такого механізму сприйняття інформації людиною і безпосередньо особливостей функціонування нейронів і картини їх зв'язків дозволили створити математичні моделі, ключовим елементом яких виступає штучний нейрон як імітаційна модель нервової клітини мозку – біологічного нейрона. Нейронна мережа є окремим випадком методів розпізнавання образів, методів кластеризації тощо, з погляду штучного інтелекту. Штучна нейронна мережа є основним напрямком в структурному підході до вивчення можливості побудови (моделювання) природного інтелекту за допомогою комп'ютерних алгоритмів. Більш того, виходячи із специфіки поданого дослідження, маємо зазначити доцільність використання штучних нейронних мереж для розпізнавання тексту, і рекламного тексту зокрема.

Нейронні мережі не програмуються, вони навчаються. Можливість навчання є однією з головних переваг нейронних мереж перед традиційними алгоритмами. З технічної точки зору, навчання полягає в знаходженні коефіцієнтів зв'язків між нейронами. В процесі навчання нейронна мережа здатна виявляти складні залежності між вхідними і вихідними даними, а також виконувати узагальнення. Це означає, що, у разі успішного навчання, мережа зможе отримати правильний результат для даних, які були відсутні в первинній вибірці, на даних якої починається навчання.

Нейронна мережа за типом навчанням без вчителя, що виконує завдання візуалізації і кластеризації є Карта Кохонена (англ. self-organizing map (SOM) або self-organizing feature map (SOFM)) [1]. Метод проектування багатовимірного простору в простір нижчого порядку застосовується також для вирішення завдань моделювання, прогнозування тощо.

Карта або мережа Кохонена (рис. 1) складається з компонентів, які називаються вузлами або нейронами. Їх кількість задається програмістом, аналітиком. Кожний з вузлів описується двома векторами. Перший – вектор ваги, що має такий самий розмір, як і вхідні дані, другий – координати вузла на карті. Опис карти відбувається з вищого вхідного вузла, до нижчого. Відповідно до відомого розміру вхідних даних, певним чином будується первинний варіант карти. В процесі навчання вектори ваги вузлів наближаються до вхідних даних. Для кожного спостереження обирається найбільш схожий по вектору ваги вузол, і значення його наближається до спостереження. Також до спостереження наближаються вектори ваги декількох вузлів, розташованих поряд. Таким чином, якщо в масиві вхідних даних два елементи були подібні, на карті їм будуть відповідати близькі нейрони активності.

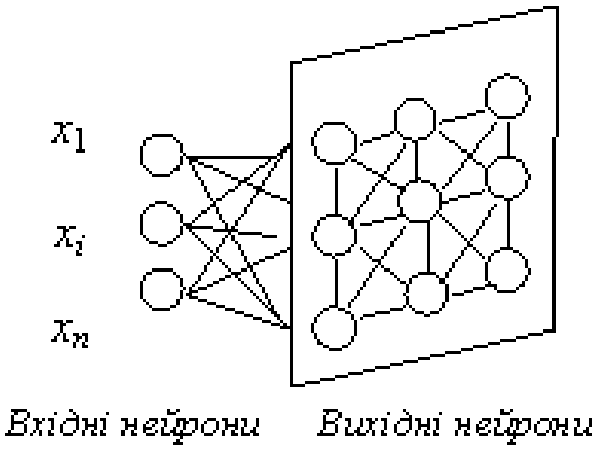


Рис. 1. Нейронна мережа Кохонена

Мета навчання в мережі Кохонена – зробити так, щоб на виході різні частини мережі однаково реагували на певні вхідні дані [2]. Навчання починається із задавання випадкових значень матриці зв'язків W_n^m . Надалі відбувається процес самоорганізації, що полягає в модифікації ваги при пред'явленні на вхід векторів навчальної вибірки. Для кожного нейрона можна визначити його відстань до вектора входу:

$$d_m = \sum_{i=1}^{N} (x_i(t) - W_i^m(t))^2 .$$

Надалі обирається нейрон *m=m**, для якого ця відстань мінімальна. На даному етапі навчання t будуть модифікуватися тільки ваги нейронів в найближчі до нейрона *m**:

$$W_n^m(t+1) = W_n^m(t) + \eta(x_n(t) - W_n^m .$$

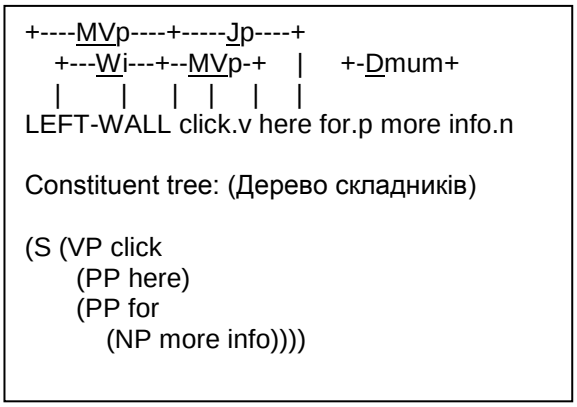
Спочатку найближчими до будь-якого з нейронів знаходяться всі нейрони мережі, надалі це наближення звужується. В кінці етапу навчання підстроюються тільки ваги найближчого нейрона. Темп навчання h(t) <1 з часом також зменшується. Розпізнання образів вибірки навчання подається послідовно, і кожного разу відбувається зважування.

Кожен нейрон несе інформацію про кластер - згусток в просторі вхідних образів, формуючи для даної групи збірний образ. Таким чином нейронна мережа Кохонена має здатність до узагальнення. Конкретному кластеру може відповідати і декілька нейронів з близькими значеннями векторів ваги, тому похибки у роботі одного нейрона не такі критичні для функціонування карти Кохонена.

Розглянемо алгоритм побудови нейронної мережі Кохонена для аналізу рекламних текстів.

На першому підготовчому етапі необхідним є процес верифікації, нормалізації речень, як структурних одиниць тексту реклами. З цією метою

необхідним є синтаксичний аналіз речень рекламного тексту на основі граматики зв'язків (Link Grammar) [3]. За допомогою граматики зв'язків можливим є синтаксичний аналіз речення, на підставі якого вибудовується дерево залежностей між парами синтаксично значущих слів, пунктуація та написання слів при цьому не зберігаються. Більш того, паралельно відбувається морфологічний аналіз, наприклад: "click here for more info" (рис. 2):



MVp – MV-зв'язок поєднує дієслова (та прикметники) із означальними фразами;
Jp – J- зв'язок поєднує прийменники із їх об'єктами;
WI – W-зв'язок використовується для поєднання головної частини речення із початком речення;
Dmum - D-зв'язок визначає зв'язок іменників із їх означеннями, означальними конструкціями

Рис. 2. Синтаксичний розбір речення парсером Link Grammar Parser

Наступним етапом підготовки тексту є його оцифрування, переведення у цифровий формат, адекватний, "зрозумілий" для нейронної мережі Кохонена. Розв'язання поданого завдання можливе двома способами використання ресурсу WordNet або використання хеш-функції.

Хешування (англ. hashing) – перетворення вхідного масиву даних довільної довжини у вихідний бітовий рядок фіксованої довжини. Такі перетворення також називаються хеш-функціями або функціями згортки, а їх результати називають хешем, хеш-кодом або дайджестом повідомлення (англ. message digest). Існує безліч алгоритмів хешування з різними характеристиками (розрядність, обчислювальна складність тощо). Вибір тієї або іншої хеш-функції визначається специфікою вирішуваної задачі. Простими прикладами хеш-функцій можуть служити контрольна сума. В загальному випадку однозначної відповідності між початковими даними і хеш-кодом немає. Тому існує безліч масивів даних, що дають однакові хеш-коди, – так звані колізії. Вірогідність виникнення колізій грає важливу роль в оцінці «якості» хеш-функккція та хешування.

Ідея використання ресурсу WordNet полягає у побудові цифрового коду для певної лексичної одиниці в залежності від відстані слів в WordNet. Нейронна мережа вимагає виконання умови про можливість порівняння значень, інакше групування нейронів для вхідного сигналу певного типу втрачає сенс. Теоретично, якщо назначити вузлу графа певний номер (випадковий), тоді спорідненим словам (синонімам, паронімам) можна назначити номери, відмінні, наприклад на +1 для синонімів, та на +10 для паронімів. Отже при проходженні

графа залежностей семантично-подібні слова будуть мати менші відстані, ніж, наприклад, антиномічні. Варіантом використання WordNet також є ідея, підґрунтям якої є сутність гіпертекстової інформації: словникові статті включаючи синонімічні рядки для кожної лексеми пропонують як посилання на статтю, при чому вказуючи те значення, яке і буде синонімічним, наприклад, синонімом до лексеми "skill" є лексема "accomplishment" у 3 або 6 значеннях, таким чином, можна зробити припущення, про те, що номер словникової статті (безпосередній номер значення для синоніму) можна співвіднести із віддаленістю синоніму від «кореневого» значення. Так, у WordNet список синонімів сортирується по частоті використання.

Далі, на підставі проведеної нормалізації речень, із подальшим їх синтаксичним розбором і оцифруванням, подаємо масив речень на запуск програми Кохонена з метою навчання нейронної мережі, використання навченої нейронної мережі для категоризації речення як рекламного/ нерекламного. Для подачі кодованої інформації на гексаграфічну нейронну мережу Кохонена створено програму на мові програмування Ruby для відстеження процесу обробки та інтерпретації результатів на виході із нейронної мережі. Так, в активному вікні Source Ads в графі Train and Visualize Neural Network (рис. 3, 4) ми маємо можливість вводу речень, натискаючи Train and Visualize Neural Network та переходячи до вкладки Digitized Ads (рис. 5, 6) ми маємо можливість відстежити результат перевірки речення на його рекламний характер.

Візуалізована нейронна мережа являє собою зображення вузлів, кольорове рішення (діапазон від чорного до білого із відтінками сірого) підпорядковане меті наочного представлення результатів. Таким чином маємо область найінтенсивнішого кольору (чорний колір), яка є ідентифікатором рекламного повідомлення (шкала насиченості кольору 100%), показник шкали насиченості кольору зменшується відповідно до виявлення нерекламного характеру речення.

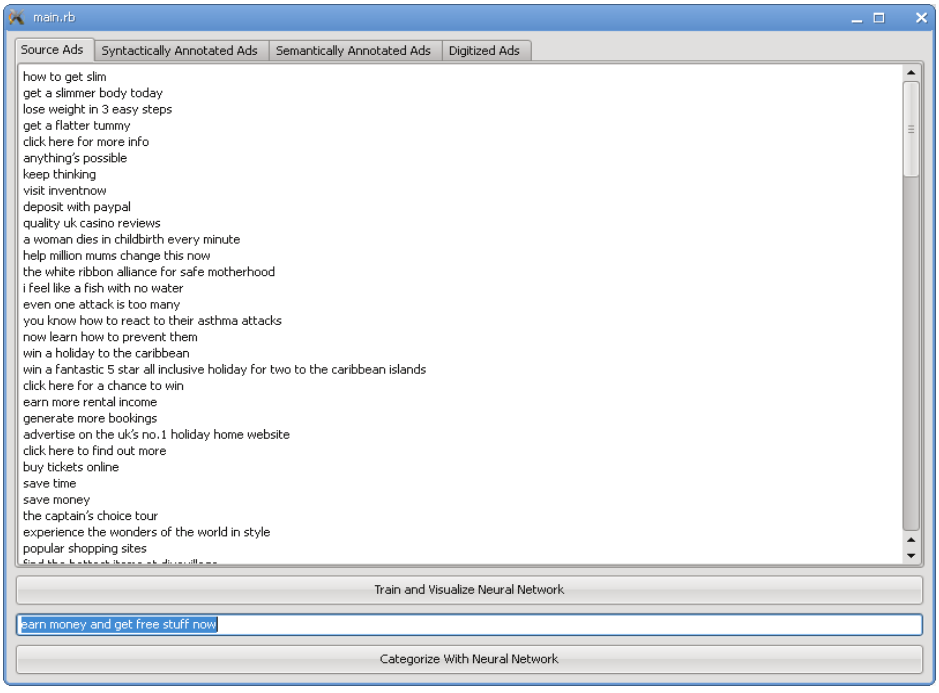


Рис. 3. Вікно вводу. Використання нейронної мережі Кохонена для категоризації речення earn money and get free stuff now

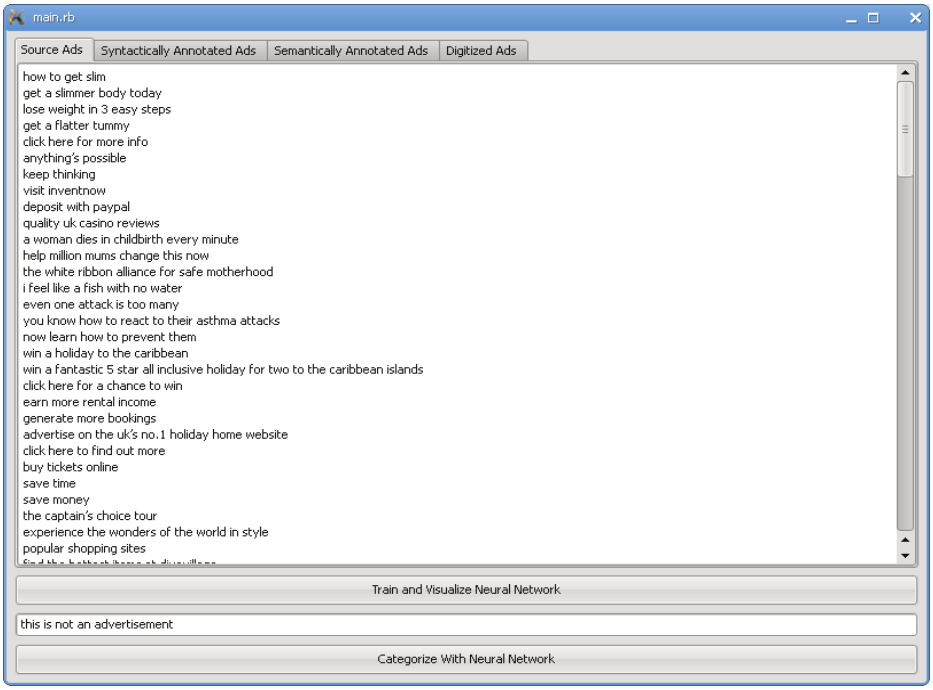


Рис. 4. Вікно вводу. Використання нейронної мережі Кохонена для категоризації речення this is not an advertisement

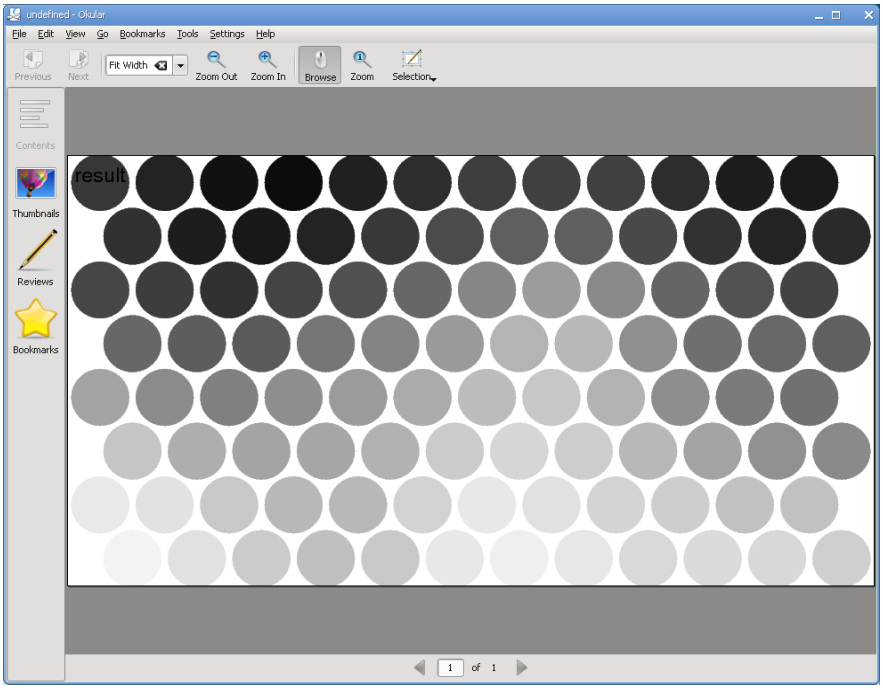


Рис. 5. Вікно виводу. Використання нейронної мережі Кохонена для категоризації речення earn money and get free stuff now result ≈ an ad

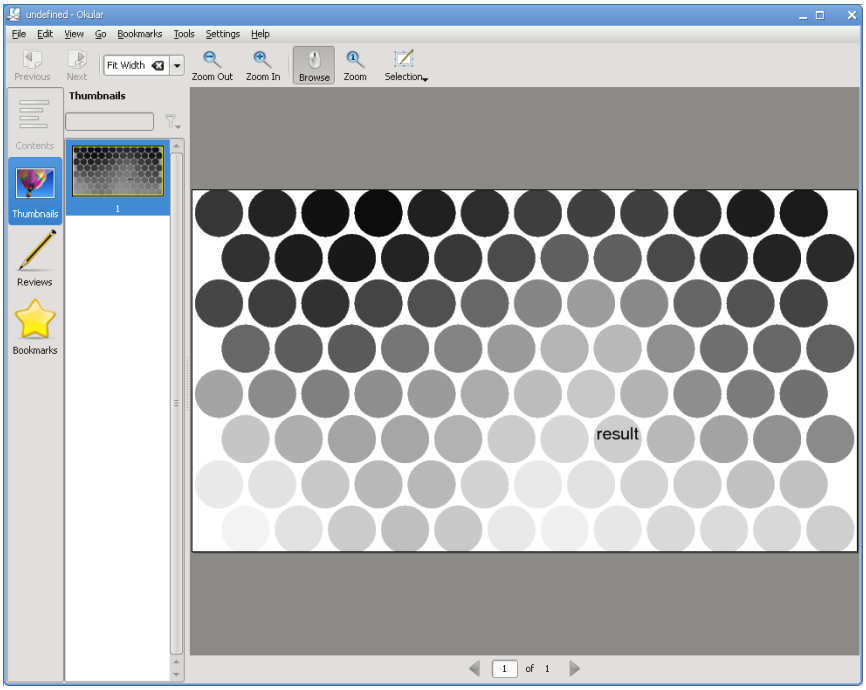


Рис 6. Вікно виводу. Використання нейронної мережі Кохонена для категоризації речення this is not an advertisment result = not an ad

Відповідно до необхідності ідентифікації не лише окремих речень як рекламних, а загалом тексту, нами пропонується два варіанти рішення цього питання. Враховуючи досить обмежений структурний потенціал рекламних текстів, а саме середню кількість речень у тексті і кількість лексем у реченні відповідно, маємо безперечно враховувати коефіцієнти насиченості кольору для всіх речень рекламного тексту. Отже наступний алгоритм дій є таким.

Визначати середній показник коефіцієнту насиченості кольору:

$$x = \sum_{i-1}^{n} \frac{x_i}{n},$$

де x_i – значення змінної X із номером, n – об'єм вибірки.

Виходячи з того, що має бути у рекламному повідомленні, відокремлюють певні структурні елементи, функціональне та змістовне навантаження яких варіюється та на підставі когнітивних особливостей сприйняття інформації людиною, маємо на меті запропонувати введення ваги для визначених показників насиченості кольору:

$$x = w_1 * a + w_2 * b + w_3 * c + w_4 * d,$$

де w_1= 0,6 , w_2= 0,4 , w_3= 0,4 , w_4= 0,8 – ваги слів у реченні, a,b,c,d – показники насиченості кольору для речень 1,2,3,4 відповідно.

Висновок і перспективи подальших досліджень у даному напрямку. В роботі покладено початок використання комбінації статистичних (нейронні мережі) і лінгвістичних методів розпізнавання рекламних текстів. Розроблена нейронна мережа і проведено навчання мережі з використанням лінгвістичних методів (граматик зв'язків і семантичної мережі WordNet). Розроблений програмний аналізатор на мові програмування Ruby навіть за умови обмеженого навчання нейронної мережі правильно ідентифікував 74% рекламних текстів. З цього можна зробити висновок про перспективність продовження розробки аналізатору і доведення якості ідентифікації до 90-95%.

ЛІТЕРАТУРА

1. Kohonen T. Self-organized formation of topologically correct feature maps / T. Kohonen //Biological Cybernetics, №43. – 1982. – pp.59-69.
2. Kohonen T. Self-Organization and Associative Memory / T. Kohonen. Berlin: Springer-Verlag. – 2001.
3. Sleator D.D. Parsing English with a Link Grammar / D.D. Sleator, D. Temperley. – Pittsburg: School of Computer Science, Carnegie Melon University. – 1993. – 14p.

Стаття надійшла до редакції 18.11.2010 р.