Рисунок 1 – Структура поділу веб-сторінки для побудови та аналізу теплової карти прокрутки

При відвідуванні сторінки, перед користувачем у більшості випадків постає та частина документу, яка вміщається на екран. Для того, щоб перейти до решти частини веб-сторінки відвідувачам необхідно прокрутити її до низу. При навігації вглиб сторінки, згідно запропонованої структури поділу, збираються дані по кількості входжень в кожен із фрагментів сторінки та середнього часу перебування у них. Зібравши ці дані та провівши їх аналіз і побудувавши теплову карту прокрутки, можна зробити висновки щодо ефективності дизайну сайту, його будови, розміщення контенту та реклами, адже все це безпосередньо впливає на привабливість ресурсу, його відвідуваність.

Висновок

В результаті дослідження впливу застосування теплових карт на відвідуваність сайту було запропоновано метод нагромадження та аналізу даних відвідування за допомогою побудови карти прокрутки веб-сторінки, що дозволило дати рекомендації щодо проектування сайтів з ефективним дизайном.

Список використаних джерел

1. Раскрутка. Секреты эффективного продвижения сайтов. / [Бабаев А., Евдокимов Н., Боде М., Костин Е., Штарев А.] — СПб.: Питер, 2013. —272 с.

УДК 004.492.3

ВИКОРИСТАННЯ SVM ТА ECOC ДЛЯ МУЛЬТИКЛАСИФІКАЦІЇ ТВОРІВ ЗА АВТОРСТВОМ

Петрушко П.П.

Київський національний університет імені Тараса Шевченка, студент

I. Вступ

Сутність проблеми встановлення авторства літературних текстів - визначити, хто є автором твору за деякими характеристиками тексту. Це буває потрібно при визначенні плагіату, знаходженні різних псевдонімів одного автора. Трапляються й анонімні твори, авторство яких можна визначити, порівнявши їх з іншими текстами, автор яких відомий. Дослідження текстів може застосовуватися в літературознавстві, історіографії, криміналістиці, захисті авторського права та інших галузях. Особливо зараз, в епоху бурхливого розвитку інформаційних систем, коли кількість різних інформаційних джерел дедалі зростає, мережа Інтернет стає все популярнішою і популярнішою, а більшість текстів уже зберігається в електронному форматі, тема визначення авторства тексту за допомогою комп'ютеризованих систем є як ніколи актуальною.

II. Постановка задачі

Припустимо у нас є деякий твір, авторство якого невідоме. Також у нас є вибірка текстів, авторів яких, ми знаємо. Задача полягає у визначенні автора, який з найбільшою ймовірністю є творцем даного тексту.

III. Особливості розв'язання задачі

У даній роботі розглядається метод визначення авторства текстів за допомогою аналізу послідовностей літер та звуків. Проаналізовано якості ймовірнісної та косинусної міри близькості. За ознаку авторства використовуються збіги двох приголосних чи голосних на стиках слів.

Даний підхід базується на представленні тексту у вигляді поліграмної моделі. У поліграмній моделі зі ступенем *n* та основою *M* текст представляється вектором $\{f_i\}, i = 1..M^n$, де f_i – частота появи *i*-ої *n-грами* в тексті. *N-грама* є послідовністю *n* символів виду: $a_1,...,a_{n-1},a_n$, причому символи a_i належать алфавіту, розмірність якого збігається з *M*. Використання таких *n*-грам дозволяє проаналізувати звукову організацію мови. Художня мова автора впорядковує ці повтори, використовуючи їх для впливу на читача, що є характерною ознакою його стилю. В даному підході ми будемо використовувати біграми(при n = 2).

Через те, що метод опорних векторів (англ. SVM, support vector machine) показав дуже хороші результати вже в багатьох великих проектах, я вирішив використовувати саме його для класифікації творів. Для прикладу, модель на основі нейронних мереж дала на порядок гірші результати ніж модель на основі SVM. Так як SVM зазвичай використовується для розв'язування задач бінарної класифікації, я за допомогою вихідного коду корекції помилок (англ. ECOC, error correction output code) скомбінував декілька бінарних SVM класифікаторів для вирішення поставленої задачі мультикласифікації.

Неформально роботу алгоритму навчання SVM можна описати в такий спосіб. Алгоритм навчання знаходить серед елементів навчальної множини точки, що лежать на кордоні двох підмножин (позитивного і негативного) і будує між цими точками поверхню, яка їх розділяє. У термінах SVM такі точки називаються опорними векторами.

SVM, крім усього іншого, має цікаву особливість, це система модульна, і вона містить в собі так звану функцію ядра, замінюючи яку можна змінювати характеристики класифікатора.

Спочатку SVM це звичайний лінійний класифікатор, тобто він може розв'язувати тільки лінійно роздільні задачі. Однак, застосовуючи нелінійне ядро, можна відобразити вихідні дані в простір більшої розмірності, де вже може існувати оптимальна гіперплощина, яка розділяє дві множини точок.

На основі проведеного аналізу творів були сформовані навчальна та тестова вибірки. В цих вибірках були використані твори десятьох авторів.

Дані в SVM подавалися у вигляді векторів:

<sign, feature : value, feature : value … >, (1)

де sign – набуває значень 1 або 0, що значить належність чи неналежність даного твору до даного автора відповідно;

feature – біграми, які були виділені в тексті;

value – частота зустрічання даної біграми в конкретному тексті.

Середня довжина вектора – 900 (кількість біграм), кількість векторів – 65.

Будемо використовувати ECOC для розбиття великої задачі мультикласифікації на менші задачі бінарної класифікації, а потім комбінування їх разом. ECOC дуже часто використовується в задачах класифікації для різних завдань зв'язаних з лінгвістикою. Кожному класу буде відповідати певне кодове слово, тобто набір значень 1 чи 0, які кожен класифікатор має видати для даного типу. Для побудови такої матриці існує дві проблеми: дизайн матриці та Хемінгова відстань між словами.

Відстанню Хемінга d(x,y) між двома двійковими послідовностями x та y (векторами) будемо вважати число позицій, в яких вони різні.

Існують різні способи побудови матиць - Identity matrix, Exhaustive matrix, Random matrix, BCH code matrix. Кожен з них використовується для різної кількості класів та забезпечує різну відстань Хемінга.

В нашому випадку будемо використовувати Exhaustive Codes для формування матриці кодових слів. Згідно з цим методом, ми будуємо кодові слова довжини $2^{k-1}-1$ для k класів. Перший рядок матриці заповнюється одиницями. Наступний заповнюється 2^{k-2} нулями, після яких іде $2^{k-2}-1$ одиниця. Тобто, для i – ого рядка 2^{k-i} нулів чергуються з 2^{k-i} одиницями. Для 10 класів, ми отримаємо матрицю з 511 класифікаторів, що є явно неефективно, так як навчити їх буду дуже енерго і часозатратним процесом. Тому оберемо з нашої матриці мінімальну кількість стовпчиків так, щоб відстань Хемінга між рядками була доволі великою. Застосувавши нескладний алгоритм, ми зменшили кількість класифікаторів до 39, з відстанню Хемінга – 16. Така модель є найбільш оптимальною для нас, тобто дана модель має можливість виправляти до 7 біт і нам потрібно буде витрачати відносно мало часу для її навчання. Це одночасно дасть нам досить високу точність, враховуючи, що бінарний класифікатор SVM показує хороші результати в лінгвістичних задачах.

Висновки

В процесі навчання було сформовано 310 моделей, які потім використовувались для класифікації на тестовій вибірці. При використанні нашої моделі на тестовій вибірці, тобто вибірці, якої модель раніше не бачила, отримано результати з точністю >=85%, тобто процент помилки становив менше ніж 15%, що є краще ніж результати отримані за допомогою моделі на основі нейронних мереж (помилка <= 20%).

Список використаних джерел

1. Тарануха В.Ю., Порхун О. В. "Автоматичне встановлення авторства текстів з використанням аналізу звукової організації мови" // Вісник Київського національного університету імені Тараса Шевченка. Серія фізико-математичні науки, 2011, №.1. – С. 63-69.
2. Thomas G. Dietterich, Ghulum Bakiri "Solving Multiclass Learning Problems via Error-Correcting Output Codes".
3. Simon Neural networks. A Comprehensive foundation. Second edition // McMaster University, Hamilton. – P. 340 – 373.

УДК 7.74: 01

ОСНОВНІ ТРЕНДИ 2014 РОКУ У ВЕБ-ДИЗАЙНІ

Решетньов І.С.

Житомирський державний університет імені Івана Франка, студент

З часів початку ери Web 2.0 сфера веб-дизайну розвивається зі швидкістю геометричної прогресії, зазнаючи при цьому вплив різних тенденцій. Було б невірним зараховувати тенденції у веб-дизайні до моди, бо причини їх появи і роль, яку вони відіграють в IT сфері, більш значущі. Власне кажучи, зараз є як би дві складові які диктують «моду» у веб-дизайні. Одна – це тренди, що диктуються розвитком сучасних мобільних технологій, каналів зв'язку і всілякого софтвера. Дані тренди носять навіть не рекомендаційний, а обов'язковий характер. В першу чергу це стосується адаптивного або реагуючого дизайну сайту. Ну і друга складова – це ті самі модні переваги, комбінація з яких дозволяє дизайнерам створювати унікальні і неповторні речі.

Технічний бум всіляких мобільних і не тільки пристроїв жорстко закликає веб-дизайнерів робити сайти, адаптовані під будь-яке розширення екрану, будь то невеликий екран кишенькового смартфона, середній розмір планшетника або величезна діагональ сучасного монітора/телевізора. Саме це призвело до того, що «адаптивність» дизайну стало чи не головною складовою при його створенні. Наступний тренд є витікаючим з першого. Якщо в тебе є великий громіздкий дизайн, то адаптувати його під велику кількість пристроїв дуже важко, тому дизайнери почали все частіше використовувати «Плоский дизайн» (Flat UI) - це ультра мінімалістичний підхід до вмісту, коли позбавляються абсолютно від усього зайвого (обсяги, шуми, градієнти і інші прикрашення), залишаючи тільки плоскі кольору і інтуїтивно зрозумілі візуальні елементи. Загалом, тільки суть і нічого зайвого.

Якщо ж говорити про «модні переваги» дизайну, які зараз з'являються все частіше, то треба відмітити вертикальний скролінг та фіксовану навігацію. Вертикальний скролінг породили особливості технічних пристроїв, коли оптимально прокручувати контент рухом знизу вгору по екрану. Варіації з посторінковим вертикальним скролінгом при наявності фіксованої навігації, суть якої в тому, що при прокручуванні сторінки зверху жорстко фіксується верхнє меню навігації виглядають дуже стильно і сучасно.

Щодо типографіки, то вона міцно влаштувалася серед сучасних трендів веб-дизайну. Якщо раніше акцент робився на красу та різноманітність вишуканих шрифтів, то тепер додалися ще й експерименти з геометрією, коли текст або слово вписуються в простір дизайну і безпосередньо взаємодіють з іншими елементами.

Також важливою «фішкою», яку можна зустріти все частіше є живі анімовані картинки. Такі картинки називаються сінемаграф. Це фотографія, на якій відбуваються незначні повторювані рухи. Сінемаграфи, які зазвичай представлені в gif форматі, створюють глядачеві ілюзію перегляду відео. Зазвичай їх отримують шляхом створення серії фотографій або відеозаписи з подальшою обробкою в графічному редакторі. При правильній подачі це виглядає дуже незвично та красиво. А вже емоційний вплив при знайомстві з сайтом може стати головним вирішальним фактором і зробити його незабутнім, так як має сильний презентаційний ефект.