[3] Image Thresholding [Електронний ресурс]. Доступно: https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_imgproc/py_thresholding/py_thresholding.html

[4] Background Subtraction [Електронний ресурс]. Доступно: https://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_video/py_bg_subtraction/py_bg_subtraction.html

УДК 004.62.77

СТВОРЕННЯ РЕЛЕВАНТНОГО КОНТЕНТУ ДЛЯ АНАЛІЗУ ПОТРЕБ СПОЖИВАЧІВ

Циганок О. П., Пронін С.В.

Харківський національний автомобільно-дорожній університет, Харків

Особливості створення релевантного контенту. Сьогодні клієнт може контактувати з компанією в різних каналах - через мобільні пристрої, соціальні платформи, рекламні банери, магазини, телевізор і т.д., Це приводить до ситуації, що відслідковувати шлях клієнта стає все складніше. Для рішення цієї задачі на сьогодні існують такі інструменти як Web Mining, Big data, методи машинного навчання. Вони дозволяють зрозуміти, що призводить до продажу і які канали неефективні. Це допомагає компаніям оптимізувати комунікацію і конверсію в різних каналах. Споживачі використовують канали по-різному, і компанії повинні враховувати це. Наприклад, деякі сегменти клієнтів вважають за краще отримувати інформацію про продукт з блогів перед відвідуванням сайту. Тому головна сторінка сайту повинна містити релевантні пропозиції та посилання на додаткову інформацію. [1]

У якості підходу до вирішення задачі пропонується застосування сегментації клієнтів по самому різному набору параметрів.

Для того щоб була можливість застосувати метод сегментації необхідно провести збір та обробку первинної інформації. Для цього скористаємося

технологіями Web Mining такими як аналіз використання веб-ресурсів, витяг веб-структур і витяг веб-контенту [2]

Аналіз використання веб-ресурсів заснований на отриманні даних з логів веб-серверів. Метою аналізу є виявлення переваг відвідувачів при використанні тих чи інших ресурсів мережі Інтернет.

Витяг веб-структур розглядає взаємозв'язку між веб-сторінками, грунтуючись на зв'язках між ними. Побудовані моделі можуть бути використані для категоризації веб-ресурсів, пошуку схожих і розпізнавання авторських сайтів. Залежно від поставленого завдання структура сайту моделюється за певним рівнем деталізації. У найпростішому випадку гіперпосилання представляють у вигляді спрямованого графа:

$$G = (D, L)$$

D - це набір сторінок, вузлів або документів; L - набір посилань. Витяг веб-структур може бути використано як підготовчий етап для вилучення веб-контенту.

Витяг веб-контенту заснований на поєднанні можливостей інформаційного пошуку, машинного навчання та Data Mining.

Аналізується зміст документів: знаходяться схожі за змістом слова та їх кількість. Потім вирішується завдання кластеризації та класифікації. Так документи групуються за смисловим близькості. Взагалі в сфері бізнес-аналітиці Web Mining вирішує наступні завдання [2]:

- опис відвідувачів сайту (кластеризація, класифікація);
- опис відвідувачів, які здійснюють покупки в інтернет-магазині (кластеризація, класифікація);
- визначення типових сесій і навігаційних шляхів користувачів сайту (пошук популярних наборів, асоціативних правил);
- визначення груп або сегментів відвідувачів (кластеризація); знаходження залежностей при користуванні послугами сайту (пошук асоціативних правил).

Створення набору даних для аналізу. Після збору даних виникне

необхідність для структурування даних для їх подальшого використання. Така структура даних називається датасет.

Одним з ефективніших інструментів для роботи з датасетом використовується на сьогодні є фреймворки Apache Spark [3].

Apache Spark вдає із себе фреймворк з відкритим вихідним кодом для реалізації розподіленої обробки неструктурованих і слабоструктурованих даних. Має підтримку таких мов як Java, Scala, Python, R .. Хоча робочими мовами є Scala і Python, згодом в нього була додана значна частина коду на Java. Це пояснюється широким поширенням странения мови Java і великою кількістю написаних на ньому проектів. Це дає можливість використовувати Spark в проектах, написаних на Java.

Фреймворк Spark включає в себе слідуючі розширення:

- Spark SQL - SQL-запит над даними,
- Spark Streaming -надстройка для обра-лення потокових даних,
- Spark MLlib -Набір бібліотек машинно-го навчання;
- GraphX - розподілена обробка графів.

Основним поняттям в Spark'e є RDD (Resilient Distributed Dataset), який представляє собою набір даних (Dataset), над яким можна робити різні перетворення. За великим рахунком, початковий RDD являє собою набір «сирих» даних, які потім за допомогою функціоналу Spark перетворюються до потрібного вигляду. Результатом застосування різних операцій до початкового RDD є новий Dataset. Серед основних операцій можна виділити:

– map (function) - застосовує функцію function до кожного елементу датасета;
– filter (function) - повертає всі еле-ти датасета, яким функція function вер-нула справжнє значення;
– distinct ([numTasks]) - повертає дата-сет, який містить унікальні елементи вихідного датасета.

Методи сегментації користувачів. Проаналізуємо різні методи сегментації користувачів контенту.

В даний час існує два основних підходи до маркетингового сегментації

ринку, розроблені Виндом в 1978 р. [4].

В рамках першого підходу, іменованого "a priory" (від лат. A priori - заздалегідь), ознаки сегментування: число сегментів, їх ємність, характеристики, карта інтересів - відомі заздалегідь, без попереднього дослідження ринку. Сегментування в даному випадку є не частиною поточного дослідження, а допоміжним засобом при вирішенні інших маркетингових завдань. Кілька прикладів схем апріорної сегментації: чоловіки і жінки, молоді й літні, північні і південні регіони, VALS- і PRiZM-кластери.

В рамках другого підходу, іменованого "post hoc" (від лат. Post hoc - після цього), виходять з невизначеності ознак сегментації і сутності самих сегментів. Підхід передбачає проведення опитування. Залежно від висловленого ставлення до певної групи змінних респонденти ставляться до відповідного сегменту. Цей підхід застосовують для споживчих ринків, сегментна структура яких не визначена у відношенні товару або наданої послуги. У нашому випадку доцільніше використовувати підхід "post hoc" так як початкова інформація не відома. Процедуру опитування можливо замінити збором інформації таким як куки, URL та метрики для визначення того, яким каналом і чому користується споживач.

При використанні підходу "post hoc" використовуються методи математичної статистики та машинного навчання такі як [4]:

Кластерний аналіз (від лат. Cluster analysis) - це математична процедура багатовимірного аналізу, що дозволяє на основі безлічі показників, що характеризують ряд об'єктів, згрупувати їх в класи (кластери) таким чином, щоб об'єкти, що входять в один клас, були більш однорідними, подібними по порівняно з об'єктами, що входять в інші класи.

Метод CART (від англ. Classification and Regression Trees - "дерева класифікації і регресії"), так само як метод CHAID, відноситься до методів "дерева класифікацій". На їх основі респондентів ділять на групи, потім кожну групу на підгрупи, що базуються на відносинах між базовими змінними сегментації і деякої залежної змінної. Зазвичай залежна змінна - це ключовий

індикатор, такий, як рівень використання блага, намір покупки і т.д. CHAID - це один з найбільш часто використовуваних методів "дерева класифікацій", але він не може впоратися з тривало залежними змінними, тому іноді використовується і CHAID, і CART, які здатні обробляти неметричні і неордінальние дані. Методи "дерева класифікацій", розділяючи респондентів, на відміну від кластерного аналізу, створюють справжні сегменти, які базуються лише на одній залежною змінною.

Нейронні мережі (Artificial Neural Network). Для їх позначення широко використовується англійська абревіатура - ANN. Штучні нейронні мережі застосовуються в багатьох сферах: при прогнозуванні економічної і фінансової діяльності, в системах обробки зображень, сигналів і т.д.

Штучні нейромережі можуть служити засобом для сегментації. Так, наприклад, самоорганізована ANN (часто користуються терміном "архітектура Кохоненом") намагається згрупувати респондентів, базуючись на їх схожих елементах. Вона відрізняється від кластерного аналізу своєю здатністю ігнорувати "шуми" в даних і тим, що нетипові індивідууми менше впливають на сегментацію. Кожна успішна ітерація робить їх внесок все менше, так що розрахунки швидко стабілізуються, ігноруючи нечасті характеристики респондентів. ANN працює тим краще, чим більше варіація або невпевненість у відповідях респондентів.

Список використаних джерел

[1] Matsiy O. Using dynamic content to increase relevance. Вісник ХНАДУ. Харків. ХНАДУ. 2021. Випуск 92. Том 1, С. 34-38.

[2] Web Mining: основные понятия. [Он-лайн]. Доступно: https://basegroup.ru/community/articles/basic-conceptions.

[3] Spark Application Overview. [Он-лайн]. Доступно: https://docs.cloudera.com/documentation/enterprise/5-6-x/topics/cdh_ig_spark_apps.html.

[4] Маркетинг. [Он-лайн]. Доступно: https://studme.org/1716100222265/marketing/marketing.