

*Майборода Аріна Миколаївна, здобувачка вищої освіти,
КПІ ім. Ігоря Сікорського, Україна*

*Науковий керівник: Солдатова Марія Олександрівна,
доцент кафедри інформаційних систем та технологій,
КПІ ім. Ігоря Сікорського, Україна*

СЕГМЕНТУВАННЯ СПОЖИВАЧІВ ДЛЯ ЯКІСНОГО МАРКЕТИНГУ У СОЦІАЛЬНИХ МЕРЕЖАХ

Анотація. Задача та проблематика просування контенту в соціальних мережах за останні кілька років набула неабиякої популярності та важливості. Саме через це маркетинг у соціальних мережах вимагає якісного сегментації споживачів.

КЛЮЧОВІ СЛОВА: СЕГМЕНТАЦІЯ, КЛАСТЕРИЗАЦІЯ, МЕТОД ГРАДІЄНТНОГО СПУСКУ, МЕТОД ЛІКТЯ

Abstract. The task and issues of content promotion in social networks have gained considerable popularity and importance over the past few years. It is because of this that marketing in social networks requires high-quality segmentation of consumers.

KEY WORDS: SEGMENTATION, CLUSTERIZATION, GRADIENT Descent METHOD, ELBOW METHOD

Вступ. Задача та проблематика просування контенту в соціальних мережах за останні кілька років набула неабиякої популярності та важливості. З подальшим розвитком автоматизації процесів, виникає питання, чи можливо здійснювати якісну рекламну кампанію за допомогою стратегії, яка була створена, використовуючи результати прогностичних моделей та моделей кластеризації? Щоб мінімізувати втручання суб'єкта, тобто фахівця з маркетингу у соціальних мережах, в процес створення плану просування. Ми відповідаємо на це питання так.

Це дозволяє налаштувати ефективну рекламну кампанію та забезпечити максимальний результат від просування продукту або послуги. Оптимальна сегментація дозволяє підібрати цільову аудиторію, зважаючи на її поведінку та інтереси в соціальних мережах. Це забезпечує зростання продажів та популярності бренду у соціальних медіа. Аналогів, в яких є автоматизоване

просування контенту досить небагато, це пояснюється відсутністю достатньої бази знань про формування стратегії просування, які використовують прогностичні моделі та моделі кластеризації. Одним з найпопулярніших є Audiense — це програмне забезпечення для соціального медіа маркетингу, розроблене, щоб допомогти користувачам максимізувати покази своїх облікових записів в соціальних мережах, таких як Twitter, з цілею популяризації. Це дозволяє їм ефективніше взаємодіяти зі своєю аудиторією, надаючи їм інструменти для аналізу даних, зібраних із цих сайтів соціальних мереж. Також, існує досить популярна платформа StackAdapt — це рекламна платформа, яка допомагає агентствам і брендам охоплювати та шукати своїх ідеальних клієнтів за допомогою привабливих блоків нативної та відеореклами. [1]

Предмет та мета дослідження

Предметом дослідження є маркетинг у соціальних мережах та його складові. Метою

дослідження є автоматизація просування контенту у соціальних мережах для підвищення ефективності працездатності фахівця з маркетингу. В рамках даної статті ми розглянемо першу важливу частину для якісного просування контенту – створення цільових аудиторій. Для цього необхідно виконати такі задачі:

- виконати огляд існуючих методів створення цільових аудиторій;
- сформулювати математичну постановку задачі;
- здійснити порівняльний аналіз методів, які використовуються при сегментації споживачів.

Задача створення цільових аудиторій.

Цільова аудиторія – це група людей, визначених як ймовірні клієнти компанії. Вони об'єднані між собою спорідненими рисами. На цьому етапі ми будемо розподіляти користувачів на сегменти.

Вхідні дані: люди, їх вік, їх інтереси.

Вихідні дані: цільові аудиторії – сегменти розподілені за віком та інтересами людей.

Алгоритм розв'язання задачі створення цільових аудиторій зображений на рисунку 1.



Рисунок 1 – Алгоритм розв'язання задачі створення цільових аудиторій

Розглянемо більш детально постановку даного алгоритму. Цільові аудиторії формуємо за допомогою кластеризації. На вхід подається множина об'єктів. На виході отримуємо об'єкти, розміщені по кластерам. Для розв'язання даної задачі будемо використовувати метод кластеризації k-means.

Мета методу – розділити n спостережень на k кластерів, так щоб кожне спостереження належало до кластера з найближчим до нього середнім значенням. Метод базується

на мінімізації суми квадратів відстаней між кожним спостереженням та центром його кластера, тобто функції 1.

$$\sum_{i=1}^N d(x_i, m_j(x_i))^2, (1)$$

де:

- d – метрика;
- x_i – i -тий об'єкт даних;
- $m_j(x_i)$ – центр кластера, якому на j -тій ітерації приписаний елемент x_i .

Головні переваги методу k-середніх – його простота та швидкість виконання.

Особливістю даного методу є те, що на вхід має подаватись кількість кластерів, які ми визначаємо наступними методами її розв'язання: методом градієнтного спуску та методом ліктя. Потім при порівнянні даних методів, ми оберемо найкращий та за ним буде створена цільова аудиторія для майбутнього просування контенту.

Основною задачею методу кластеризації у нашому випадку є сегментація цільової аудиторії на підмножини. У нашому випадку, це категорії (кулінарія, медицина, IT, і т.д.). Тому нашою задачею є сегментація на кластери цільової аудиторії за категоріями. Оскільки в базі даних є інформація про їхні вподобання категорій, а їх може бути кілька, то сегментація буде відбуватись по найулюбленішій категорії цієї людини.

Кластеризація методом k-means з визначенням кількості кластерів методом градієнтного спуску

Градієнтний спуск – ітераційний алгоритм оптимізації, який дозволяє мінімізувати функціонал помилки шляхом оновлення параметрів шуканої функції у зворотному напрямку градієнта. [6]

Ідея градієнтного спуску: на кожному кроці оновлювати параметри шуканої функції у зворотному напрямку градієнта функціоналу помилки, тим самим здійснюючи оптимізацію у напрямку найшвидшого спуску.

Для пошуку мінімуму множини кластерів функції k-means $J(C)$ використовується метод градієнтного спуску.

Математична постановка задачі

Для вибору числа кластерів використовується значення функції k-means $J(C)$. Обирають те число кластерів, починаючи з якого значення $J(C)$ спадає не так швидко.

Введемо наступні позначення:

- $C = \{\vec{C}_1, \vec{C}_2, \dots, \vec{C}_k\}$ — множина центроїдів;
- $X = \{\vec{X}_1, \vec{X}_2, \dots, \vec{X}_n\}$ — множина даних, яку необхідно кластеризувати;

Розглянемо функцію класичного алгоритму k-means. Вона наведена у формулі 2.

$$J(C) = \sum_{k=1}^{|X|} \sum_{i=1}^n (x_{ij} - c_{k,j})^2 \quad (2)$$

Далі потрібно мінімізувати суму квадратів Евклідової відстані елементів кластера від відповідного центроїда. Представлення мінімізації наведене у формулі 3.

$$J(C) = \sum_{k=1}^{|C|} \sum_{i=1}^{|X_k|} f^2(\vec{X}_i, \vec{C}_k) \quad (3)$$

Знайдемо похідну нульової функції. Функція наведена у формулі 4. [2]

$$\frac{\partial J(C)}{\partial C_{a,b}} = \sum_{i=1}^{|X|} \frac{\partial \min_{j=1}^{|C|} f^2\{\vec{x}_i, \vec{c}_j\}}{\partial C_{a,b}} \quad (4)$$

Кластеризація методом k-means з визначенням кількості кластерів методом ліктя

При кластеризації методом k-середніх кількість кластерів найчастіше оцінюють за допомогою «методу ліктя». Він передбачає багаторазове циклічне виконання алгоритму зі збільшенням кількості кластерів, що вибираються, а також подальшим відкладанням на графіку бала кластеризації, обчисленого як функція від кількості

кластерів. Основним показником методу ліктя є $J(C)$ - сума квадратів помилок. [7]

Основна ідея методу ліктя полягає в тому, що в міру збільшення числа кластерів k розподіл вибірок буде більш точним, а ступінь агрегації кожного кластера поступово збільшуватиметься, тому квадрат помилок, природно, поступово стануть меншими. Більше того, коли k менше числа істинних кластерів, збільшення k значно збільшить ступінь агрегації кожного кластера, тому зниження $J(C)$ буде більшим, і коли k досягне числа істинних кластерів, значення, отримане шляхом збільшення k . Ступінь агрегації, повертається швидко стає менше, тому зниження $J(C)$ буде різко зменшуватися, а потім поступово згладжуватиметься, оскільки значення k продовжує збільшуватися, що означає, що відносини між $J(C)$ і k мають форму коліна, і це коліно. Відповідне значення k є кількістю справжніх кластерів у даних. Звичайно, саме тому метод називається методом ліктя. [7]

Математична постановка задачі

Математична постановка задачі полягає у мінімізації множини кластерів функції витрат k-means та наведена у формулі 6.

$$J(C) = \sum_{k=1}^K \sum_{i \in C_k} \|x_i - \mu_k\|^2 \rightarrow \min, C \quad (6)$$

Де:

C — множина кластерів потужності K

μ_k — центроїд кластера C_k

Для вирішення питання вибору кількості кластерів потрібно знайти мінімум, але оскільки такий мінімум буде при тому, що кожна точка – це кластер одного елементу, то потрібно скористатись правилом, побудувавши графік, знайти точку, де $J(C)$ падає не так швидко. Дане представлення наведене формулою 7.

$$D(k) = J(C_k) - J(C_{k+1}) \vee \frac{J(C_{k-1}) - J(C_k)}{J(C_k) - J(C_{k+1})} \rightarrow \min, C \quad (7)$$

Візьмемо частину нашої цільової аудиторії та побудуємо графік, аби визначити оптимальну кількість кластерів для такої кількості людей. Це показано на рисунку 2.

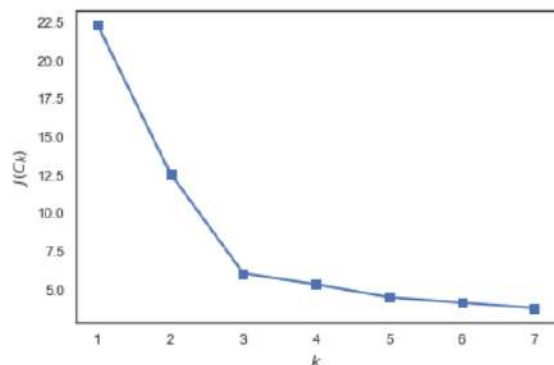


Рисунок 2 – Графік побудови методу ліктя

На рисунку бачимо, що при зміні k від 3 до 4, значення $J(C)$ падає не так швидко, тому для цього варіанту оптимальною кількістю кластерів буде 3 кластери.

Порівняння методів ліктя та градієнтного спуску

Для порівняння методів пошуку кількості кластерів проведемо експеримент. Отже, спочатку візьмемо таблицю даних можливих інтересів, для використання у подальшому обрахунку. Це показано на рисунку 3.

Results			Messages
	Id	Name	
1	1	Кулінарія	
2	2	Медицина	
3	3	ІТ	
4	4	Ігрова сфера	
5	5	Кіно	
6	6	Театр	
7	7	Фітнес та спортзал	
8	8	Крипто-сфера	
9	9	Бізнес	
10	10	Туризм	
11	11	Мистецтво	

Рисунок 3 – Таблиця інтересів

Оскільки всього у нас вікових категорій три, то за максимального порогу ми отримаємо 33 кластери. Візьмемо інформацію з бази даних та виконаємо кластеризацію з пошуком кількості кластерів методом ліктя та методом градієнтного спуску.

Результати відображені у вигляді кривих на графіках на рисунках 4 та 5.

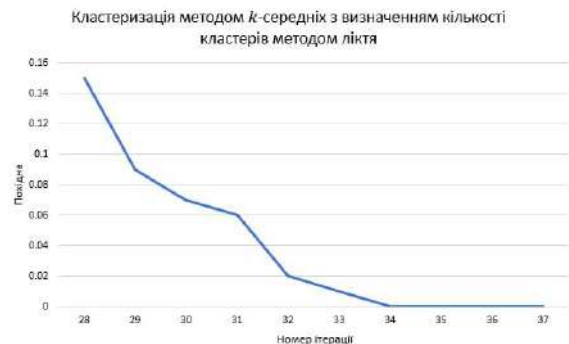


Рисунок 4 – Кластеризація методом k-середніх з визначенням кількості кластерів методом ліктя

Для методу ліктя бачимо, що починаючи з 34-ої ітерації, похідна дорівнює нулю. Але так як наша максимальна кількість кластерів 33, то обираємо 33. Якщо кожна людина належить до певної вікової категорії та має свої інтереси, то вона може бути віднесена до конкретного кластеру з цими ж характеристиками, тому людей, які не потрапили в кластер, немає. Отже, оскільки метрика Precision становить 100%, ми можемо стверджувати, що точність дорівнює 100%. Це означає, що значення метрики Recall буде дорівнювати 0.

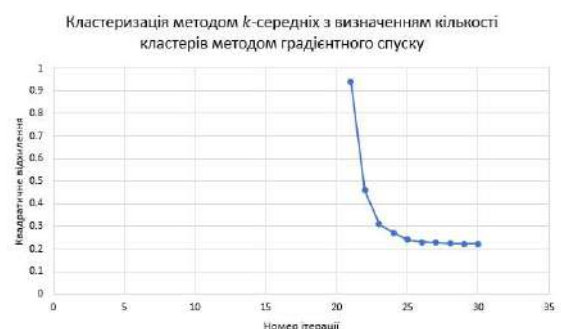


Рисунок 5 – Кластеризація методом k-середніх з визначенням кількості кластерів методом градієнтного спуску

Для методу градієнтного спуску бачимо, що починаючи з 24-ої ітерації, квадратичне відхилення падає не так швидко, як раніше. Розглянемо два варіанти, де візьмемо кількість кластерів 25 та 30.

Для 25 кластерів дей, які не потрапили в кластер, дорівнює 25% та для 30 кластерів буде дорівнювати 10%. Отже, чим більше кластерів – тим більше точність

кластеризації. Вони відрізняються Precision, у першому варіанті Precision складає 75%, а в другому 90%. Таким чином Recall дорівнює 25% та 15% відсотків відповідно.

Отже, висновок полягає в тому, що використання методу градієнтного спуску призведе до більшої повноти кластерів (більше людей буде в кожному кластері), але

може привести до втрати пріоритетності інтересів. З іншого боку, метод ліктя забезпечує найкращу точність розрахунків, що особливо важливо для кластеризації цільової аудиторії. Таким чином, обираючи метод ліктя, користувач матиме можливість обрати цільову аудиторію з найбільшим інтересом саме до цієї теми.

Висновок. В даній роботі проведено дослідження в рамках сегментації споживачів для створення якісної рекламної кампанії. Розподілення людей визначено за допомогою кластеризації методом k-середніх. Також, ми дізналися, що даний метод потребує на вхід очікувану кількість кластерів. Її ми знаходимо двома методами: методом ліктя та методом градієнтного спуску. Після чого ми обрали найкращий та прийшли до висновку, що метод ліктя дозволяє це зробити швидше та з меншою похибкою для великого об'єму даних. Описаний підхід допоможе експертам у галузі маркетингу значно полегшити процес складання плану просування контенту, створення якісної рекламної кампанії.

Список інформаційних джерел

1. «Audiense» review. – 2022. – [Електронний ресурс]. – Режим доступу до ресурсу: <https://crozdesk.com/software/audiense>.
2. Solution of the clustering problem by the method of gradient descent. 2022. – [Електронний ресурс]. – Режим доступу до ресурсу: <https://habr.com/ru/articles/188638/>.
3. Кластеризація методом k-середніх. – 2022. – [Електронний ресурс]. – Режим доступу до ресурсу:
4. https://www.wiki-data.uk-ua.nina.az/Кластеризація_методом_k-середніх.html
5. Метод ліктя. – 2022. – [Електронний ресурс]. – Режим доступу до ресурсу: <https://www.helenkapatsa.ru/mietod-loktia/>
6. Розв'язування задачі кластеризації методом градієнтного спуску. – [Електронний ресурс]. – Режим доступу до ресурсу: <https://habr.com/ru/post/188638/>
7. Градієнтний спуск. – [Електронний ресурс]. – Режим доступу до ресурсу: https://uk.wikipedia.org/wiki/%D0%93%D1%80%D0%B0%D0%B4%D1%96%D1%94%D0%BD%D1%82%D0%BD%D0%B8%D0%B9_%D1%81%D0%BF%D1%83%D1%81%D0%BA
8. Кваліфікаційна робота на тему “Дослідження методів кластеризації часових рядів за значеннями параметрів”. – Зінченко. – 2021. – Харків: Харківський національний університет радіоелектроніки.