УДК 004.85

КЛАСТЕРИЗАЦІЯ КОНТЕНТУ З ВИКОРИСТАННЯМ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ

Мартиненко А.О.
e-mail: andrii.martynenko1@nure.ua
Харківський національний університет радіоелектроніки, каф.ПІ
М. Харків, Україна

Examination of the use of machine learning for website content clustering. Clustering the semantic core allows for grouping similar keywords into distinct categories, thereby simplifying subsequent analysis and guiding the creation of relevant content. It is proposed to use K-means, Mini-batch K-means, Deep Embedded Clustering and Spectral Clustering to identify thematically similar groups of queries. The approach helps to uncover hidden structures and themes in large keyword sets, enabling more precise SEO strategies and an organized content architecture. Experimental evaluations highlight the effectiveness of each algorithm across various data volumes, noting differences in accuracy, computational demands, and interpretability.

В сучасних умовах стрімкого розвитку веб-технологій та загальної цифровізації якісна пошукова оптимізація (SEO) [1] стає одним із ключових чинників для успішного просування сайтів у пошукових системах. Конкуренція за високі позиції у видачі зростає з кожним днем, що зумовлює необхідність постійного вдосконалення методів та засобів оптимізації. Однією з фундаментальних складових оптимізації сайтів в пошукових системах є опрацювання семантичного ядра – це процес підбору та структурування ключових слів, що близькі за написанням та змістом, які відображають зміст та тематику веб-сайту за пошуковими намірами користувача.

Зазвичай семантичне ядро містить велику кількість ключових фраз, котрі можуть перетинатися за змістом, бути багатозначними або частково дублюватися. Внаслідок цього виникає проблема систематизації та ефективного групування ключових слів, що необхідно для побудови релевантних та тематичних посадкових сторінок, підготовки якісного контенту та визначення подальшої пріоритетності робіт з оптимізації сайту для успішного просування в результатах пошукової видачі.

Застосування ручного аналізу ключових запитів у таких умовах може бути надто трудомістким та містити високий ризик помилок чи пропуску важливих закономірностей.

Натомість методи машинного навчання, зокрема алгоритми кластеризації, дозволяють значно автоматизувати цей процес. Кластеризація дає змогу згрупувати велику кількість ключових слів за їх семантичною близькістю, а також виділити приховані теми чи підтеми в масиві даних.

Метою дослідження є порівняння ефективності застосування різних алгоритмів машинного навчання (K-means, Mini-batch K-means, Spectral Clustering та Deep Embedded Clustering) для кластеризації веб-контенту з метою вдосконалення SEO-стратегії та підвищення якості групування семантичного ядра сайтів.

Завдання дослідження полягає в аналізі принципів роботи алгоритмів K-means, Mini-batch K-means, Spectral Clustering та Deep Embedded Clustering, порівнянні одержаних результатів за обчислювальними витратами й часом виконання, а також визначенні найбільш доцільного алгоритму залежно від структури даних і доступних ресурсів.

Базовий алгоритм K-means [2], що часто є в основі інших алгоритмів, дозволяє групувати об’єкти за допомогою визначеної кількості кластерів. Суть алгоритму полягає у пошуку центра мас кожного кластеру та ітеративному переобчисленні їх положення доти, доки не буде досягнута збіжність.

Для опрацювання дуже великих наборів ключових фраз доцільно використовувати Mini-batch K-means, у якому обробка даних відбувається невеликими блоками. Такий підхід дає змогу суттєво пришвидшити конвергенцію алгоритму завдяки зменшенню обчислювальних витрат на кожному кроці, що особливо важливо, коли йдеться про десятки чи сотні тисяч ключових слів.

Алгоритм Spectral Clustering належить до спектральних методів, які спочатку будують матрицю суміжності або схожості між об’єктами, а потім перетворюють її у спектральний простір за допомогою власних векторів. У цьому просторі об’єкти стають лінійно роздільними, тож застосування традиційних методів K-means в основі алгоритму дозволяє успішно виявляти кластери. Такий підхід дає змогу краще враховувати нелінійні зв’язки між об’єктами та виявляти складнішу внутрішню структуру даних.

Алгоритм Deep Embedded Clustering (DEC) інтегрує спеціальну нейронну мережу для зменшення розмірності та кластеризацію за допомогою K-means. Спочатку автоенкодер навчається відображати вхідні дані, тобто текстові вектори ключових слів у простір меншої розмірності так, щоб зберегти головні особливості. Потім на стиснутих ознаках застосовується K-means для знаходження кластерів.

Дані для тестування моделей включали інформацію про сторінки веб-сайтів, таку як ключові слова. Ці дані оброблялися через TF-IDF для векторизації текстових характеристик, а поведінкові сигнали та метадані слугували основою для кластеризації. Програмні засоби, зокрема Python з бібліотеками scikit-learn та tensorflow, використовувалися для реалізації моделей. Метрики оцінки включали час виконання (середній час на одну ітерацію кластеризації), точність за допомогою Silhouette Score та Adjusted

Rand Index (ARI), а також масштабованість на наборах даних від 1000 до 100,000 сторінок. Результат наведено в таблиці 1.

Таблиця 1 – Порівняння моделей

Модель	Час	Точність	Масштабованість
K-means	2.5 с	87%	Висока
Mini-batch K-means	1.8 с	85%	Дуже висока
Spectral Clustering	4.5 с	90%	Середня
DEC	6.0 с	92%	Висока

Кластеризація контенту на основі алгоритмів машинного навчання є важливим етапом для побудови якісної стратегії просування веб-сайтів в результатах пошукової видачі. Результати порівняння показали, що кожна модель має свої сильні та слабкі сторони залежно від обраного сценарію використання. Базовий K-means забезпечує швидке й точне групування сторінок, проте його ефективність знижується на великих обсягах даних. Mini-batch K-means демонструє оптимальні результати для великих наборів завдяки зменшенню обчислювальної складності, хоча його точність трохи нижча. Spectral Clustering підходить для даних із нелінійною структурою, але потребує більше ресурсів. Найкращі результати за точністю показала Deep Embedded Clustering (DEC), яка інтегрує нейронні мережі для зменшення розмірності та обробки багатовимірних даних, що робить її ідеальною для складних завдань, хоча й потребує значних обчислювальних ресурсів.

Під час дослідження алгоритми K-means, Mini-batch K-means, Spectral Clustering та Deep Embedded Clustering демонстрували різний рівень точності та швидкодії залежно від масштабу даних та ресурсів.

У результаті проведеного дослідження встановлено, що застосування алгоритмів машинного навчання для кластеризації контенту забезпечує істотні переваги в оптимізації процесів SEO. Використання кластеризації суттєво скорочує час ручного сортування та групування великих обсягів ключових слів, що є критично важливим за умови динамічного розширення семантичного ядра.

Список використаних джерел:

1. Enge E., Spencer S., Stricchiola J. The Art of SEO: Mastering Search Engine Optimization. O'Reilly Media, 2023. 925 p.

2. Bishop C. M. Pattern Recognition and Machine Learning. Springer Science & Business Media, 2006. 758 p.