ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 28 МАТЕМАТИКИ Выпуск 1
2021

И. А. Дмитроченко, **Ю. И. Пастухова** (Москва, РТУ МИРЭА, ЦЭМИ РАН)**. Применение методов кластерного анализа для выявления сайтов категории 18+ при размещении рекламы.**

УДК 004.9

Резюме: Актуальность выбранной тематики обусловлена быстрым развитием технологий автоматизированной закупки рекламы, которая предусматривает отсутствие прямого участия человека. В связи с этим необходимо внедрение методов автоматической проверки сайта на отсутствие контента, способного нанести вред рекламируемому бренду. В данной статье будет описано использование методов кластеризации для выявления сайтов, содержащих контент для взрослых.

Ключевые слова: реклама, кластеризация, взрослый контент, python.

В настоящее время интернет-реклама является единственным сегментом рекламного бизнеса, который показывает положительную динамику роста. Так по данным Ассоциации Коммуникационных Агентств России в 2019 году бюджет рекламных размещений в интернете вырос на 20% по сравнению с предыдущим годом и составил около 50% от общего объема закупки рекламы (Ассоциация коммуникационных агентств России, 2020) ([1]). Одним из наиболее растущих направлений в интернет-рекламе является RTB-технология (от англ. real-time-bidding–«аукцион в реальном времени»). Показ рекламы происходит на множестве сайтов, входящих в рекламную сеть, при этом список сайтов не известен заранее. Поэтому рекламные сети должны выявлять сайты, которые могут нанести вред рекламодателям. Однако в России лишь несколько компаний, большинство из которых иностранные, имеют необходимые технологии для решения данной задачи.

Для построения собственного решения использовались математические методы кластеризации и их реализация на языке программирования Python. В качестве материала для кластеризации выбран набор данных, состоящий из текстового наполнения 43 сайтов, 19 из которых относятся к категории «материалов для взрослых», а 23 относятся к различным «безопасным категориям». Из текста были удалены все знаки препинания, а также стоп-слова (шумовые слова — слова, не несущие смысловой нагрузки). Затем каждое слово в тексте было нормализовано, т. е. приведено к начальной форме, при помощи морфологического анализатора pymorphy2 [2].

После этапа предобработки был выбран способ перевода слова в числовую форму с помощью статистической меры TF-IDF (от англ. termfrequency — «частота слова», inversedocumentfrequency — «обратная частота документа»). Мера TF-IDF оценивает важность слова не только в контексте документа, но и в контексте корпуса (совокупности) документов, а именно, TF указывает на частоту, с которой слово встречается в документе, IDF характеризует количество документов, содержащих слово среди всех документов корпуса. Обозначим $tf(t,d)$ относительную частоту, с которой слово t встречается в документе d:

$$tf(t,d) = \frac{n_t}{\sum_k n_k},$$

где n_t — количество вхождений слова t в документ d, $\sum_k n_k$ — общее число слов

в документе. Величина $idf(t, D)$ определяет логарифм отношения числа документов $|d|$ в корпусе D к количеству документов, в которых встречается слово t :

$$idf(t, D) = \log \frac{|d|}{|\{d \in D | t \in d\}|}.$$

Таким образом, каждому i-му слову, входящему в j-й документ ставится в соответствие величина $tf(t_i, d_j) \times idf(t_i, D)$. Вес слова, вычисленного по данной формуле, будет тем выше, чем больше его относительная частота в пределах одного документа и меньше частота документов его содержащих среди всех документов. В результате формируем матрицу размерности (n, m), где n–количество документов в корпусе, m — количество слов в корпусе [3].

Для разделения на кластеры был выбран метод k-средних, как более удобный в рассматриваемом случае разделения сайтов на два кластера: взрослый контент и приемлемый контент. Количества итераций алгоритма существенно зависит от выбора начальных центров кластера. Точки инициализации выбираются случайным образом из числа точек набора данных. Для подбора числа итераций построим «локтевой» график зависимости inertia от числа итераций:

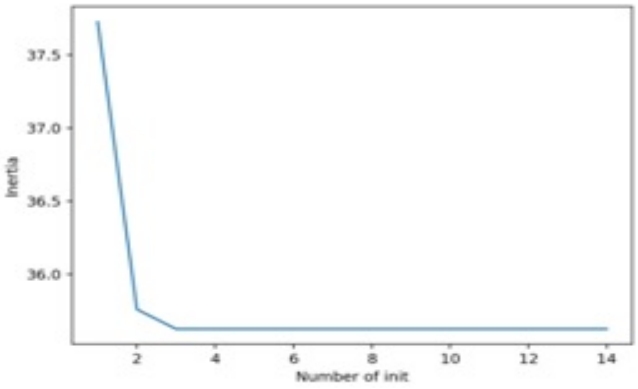

Рис. 1. График зависимости inertia от числа итераций алгоритма

Из графика следует, что алгоритм перестает изменять центры кластеров после 3 итераций. Результат работы алгоритма после понижения размерности данных представлен на рис. 2, на котором видно, что алгоритм успешно разделил датасет на две группы:

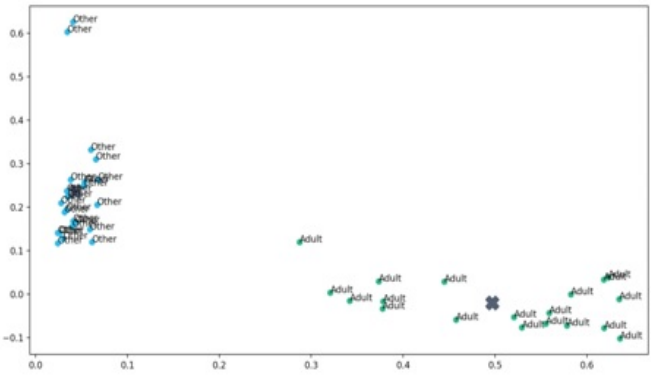

Рис. 2. Визуализация разбиения набора данных на кластеры

Для понижения размерности применялся метод главных компонент с использованием модуля skleam.decomposition, написанного на языке Python. Вычислительные эксперименты показали, что успех кластеризации был обусловлен выбором метрики TF-IDF, позволившей максимизировать расстояние между элементами двух групп. Поэтому алгоритму достаточно трех итераций до момента остановки, а выбор начальных центров будет успешен в большинстве случаев.

Созданный на языке Python алгоритм можно использовать повторно, сохранив полученный словарь TF-IDF и обученную модель в формате pkl. Словарь можно расширять и дообучать модель на новых данных, подгружая из файла.

СПИСОК ЛИТЕРАТУРЫ

1. Ассоциация коммуникационных агентств России. (март 2020 г.). Объем рекламы в средствах ее распространения в 2019 году. Получено из АКАР: https://www.akarussia.ru/knowledge/market size/id9112
2. *Korobov M.* Morphological Analyzer and Generator for Russian and Ukrainian Languages. — Analysis of Images, Social Networks and Texts, 2015, p. 320–332.
3. Scikit-learn developers. (4 August 2020 г.). Scikit-learn user guide. Release 0.23.2. Получено из Sklearn: https://scikit-learn. org/stable// downloads/scikit-learn-docs.pdf

UDC 004.9

Dmitrochenko I. A., ***Pastuchova Ju. I.*** (Moscow, RTU MIREA, CEMI RAS).

Application of cluster analysis methods to identify sites of category 18+ when placing ads

Abstract: The relevance of the selected topic is caused by a rapid development of programmatic advertising buying, which includes lack of human's direct participation. Thereby, it requires methods of automatic site screening to find content, that can be harmful for advertiser. In this article we describe how clustering methods can be used to identify sites with adult content.

Keywords: advertising, clustering, adult content, python.