DOI 10.37882/2223-2966.2021.01.21

ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА ИНТЕРЕСОВ ПОЛЬЗОВАТЕЛЕЙ

THE APPLICATION OF MACHINE LEARNING METHODS FOR THE ANALYSIS OF USER INTERESTS

V. Monastyrev
S. Molodyakov

Summary. This article discusses the process of analyzing user data to identify target interest groups. The analyzed data includes photos, text, geotags, and marks on publications ("like" or "dislike"). Machine learning methods are used to analyze information. The analyzed information is recorded in a database, where different types of information are combined by a common identifier. Based on the analyzed information, target groups are selected by applying an SQL query.

Keywords: machine learning, neural network, recommendation system, image, text tonality, rating.

Монастырев Виталий Викторович
Аспирант, Санкт-Петербургский политехнический университет Петра Великого
vit34-95@mail.ru
Молодяков Сергей Александрович
Д.т.н., профессор, Санкт-Петербургский политехнический университет Петра Великого
molodyakov_sa@spbstu.ru

Аннотация. В данной статье рассматривается процесс анализа данных пользователей для выделения целевых групп по интересам. В качестве анализируемых данных рассматриваются фото, текст, геометки и отметки на публикациях ("нравится" или "не нравится"). Для анализа информации применяются методы машинного обучения. Проанализированная информация записывается в базу данных, где различные типы информации объединяются общим идентификатором. На основе проанализированной информации выделяются целевые группы путем применения SQL-запроса.

Ключевые слова: машинное обучение, нейронная сеть, рекомендательная система, изображение, тональность текста, рейтинг.

Введение

В современном мире человечество ежесекундно оставляет в сети Интернет множество различной информации: тексты, картинки, геометки и прочее. Подобная информация может быть проанализирована с применением методов машинного обучения. При помощи алгоритмов на основе публикуемой информации можно выявить интересы пользователя, который публикует эту информацию. Выявление интересов пользователя в дальнейшем можно использовать при таргетированной рекламе: предлагать товары в зависимости от интересов. Предложения на основе интересов в свою очередь повышают CTR (click-through rate) — соотношение кликов по рекламе к числу ее показов и повышают продажи рекламируемых товаров.

Подобные системы анализа интересов существуют практически в каждом крупном сервисе. В качестве примера может послужить рекомендательная система YouTube [1]. В основе данного алгоритма лежит глубокое обучение, реализованное при помощи библиотеки TensorFlow. Система состоит из двух нейронных сетей: одна используется для генерации кандидатов и одна для ранжирования. На вход сети генерации кандидатов поступают события из истории активности пользователя, также извлекается небольшое подмножество видео из большого корпуса видеороликов. Из минусов стоит отметить, что такая система пригодна только для крупных проектов с большими объемами информации. Другой пример — система Яндекс Диско. Яндекс Диско использует множество различных подходов и строит сотни моделей, которые затем обрабатываются Матрикснетом, специальным методом машинного обучения, который ранжирует и выбирает наиболее релевантные на основе множества факторов. Из минусов стоит отметить, что подобная система пригодна для обработки больших объемов информации, но может плохо показать себя на небольшом объеме данных. Кроме того, для обучения всех моделей понадобится большое количество вычислительных ресурсов и времени.

Из рассмотренных примеров стоит отметить, что модели справляются с полученными задачами, автоматизирован процесс дообучения, обучения производилось на больших объемах информации. С другой стороны, весь код подобных систем является закрытым, они ориентированы на определенный тип данных, а для обучения моделей тратятся значительные ресурсы.

Постановка задачи

Известны рекомендательные системы, построенные на основе небольшого набора данных. Они собирают и анализируют разнородные данные. За счет сбора разнородных данных удается сократить объем анализиру-

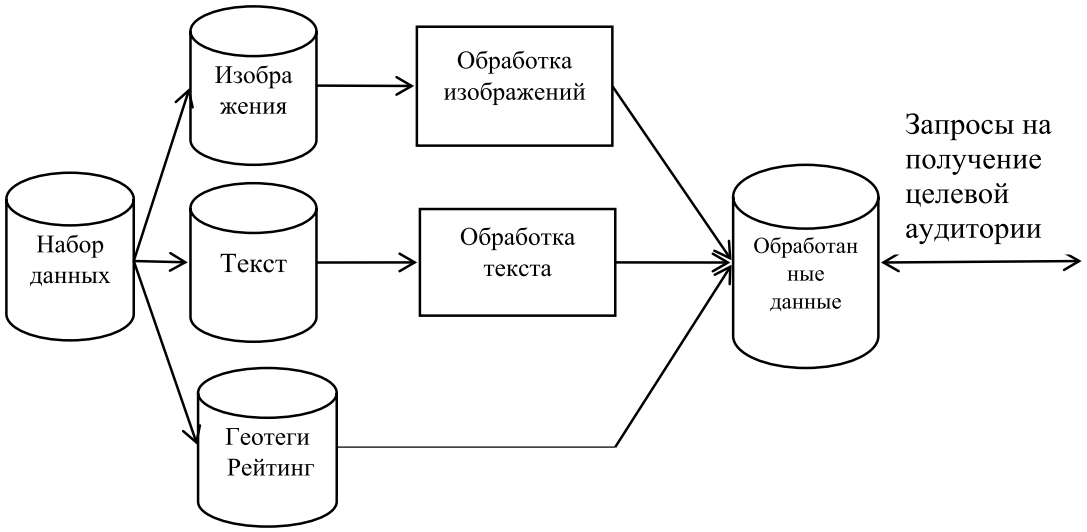

Рис. 1. Схема рекомендательной системы

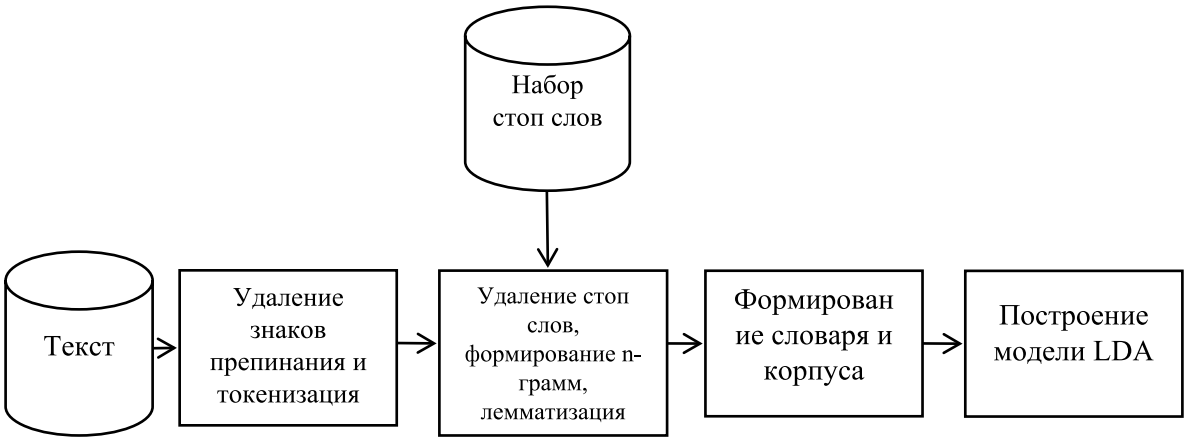

Рис. 2. Алгоритм выделения основной тематики из текста

емых данных, повысить качество рекомендаций. Большинство систем построено на основе независимого анализа разнородных данных с последующим объединением результатов анализа [1, 2].

Целью данной работы является разработка и исследование системы, позволяющей выделять группы пользователей на основе их интересов. В качестве анализируемых данных будут использоваться изображения, тексты, геометки и отметки рейтинга публикаций. Она будет работать с относительно небольшим набором данных, иметь возможность модульно добавлять новые типы данных в процесс обработки.

Архитектура системы представлена на рис. 1. Набор данных разделяется на 3 типа: изображения, текст и геотеги и рейтинг. Изображения и текст подвергаются дополнительной обработке. Обработка изображений предполагает, что мы распознаем объект на данном фото. Обработка текста предполагает, что мы распознаем тематику текста и его тональность. Геотеги и рейтинг дополнительно не обрабатываются, а используются в исходном виде для выделения аудитории на основе геолокации и понравившихся записей. После обработки вся эта информация вновь объединяется на основе уникального идентификатора и формируется целевая витрина данных. В дальнейшем данная витрина будет использоваться для получения целевой аудитории при помощи простого SQL-запроса, что позволит организовать быстрый доступ к данным [3].

Проверка и анализ разработанной системы осуществлялся на наборе данных из реальной работающей системы, который содержит: 127 фотоизображений, 122

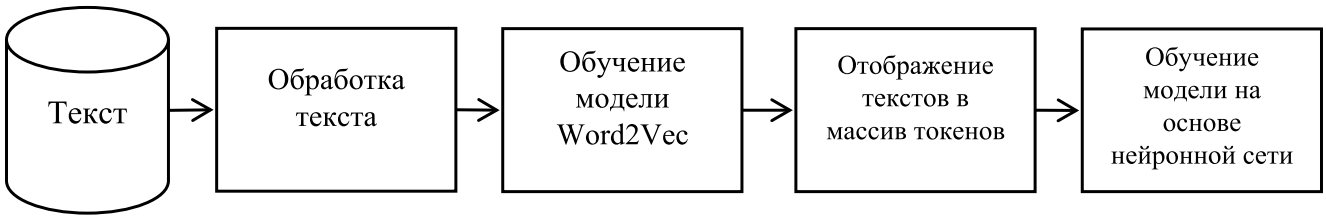

Рис. 3. Алгоритм определения тональности текста

и 180 подписей и комментариев под записями соответственно, 496 отметок рейтинга, 47 геометок.

Разработка алгоритмов анализа текстов

Для анализа текста были разработаны и применены 2 алгоритма, которые работают параллельно. Первый алгоритм определяет тематику, а второй тональность текстовых сообщений. Для выделения основной тематики из текста предлагается использовать следующую последовательность шагов:

1. Из текста удаляются знаки препинания и происходит его токенизация.
2. Из текста удаляются стоп-слова, которые агрегированы в отдельной базе данных и формируются n-граммы.
3. На основе очищенного текста формируется словарь и корпус.
4. Строится модель LDA.

Таким образом, к алгоритму LDA (латентное размещение Дирихле) [4] был добавлен шаг удаления знаков пунктуации и токенизации текста. Кроме того, так как работа происходила с русским текстом, были добавлены стоп-слова для русского языка.

Из текста удаляются лишние символы и стоп-слова, формируются словари и корпус и затем строится модель LDA [5], основанная на вероятности:

$$p(d,w) = \sum_{t \in T} p(d)p(w|t)p(t|d).$$

Для удаления знаков препинания используется библиотека Gensim. В качестве набора стоп-слов мы предлагаем использовать набор из пакета nltk.corpus, а для создания n-грам использовать библиотеку Gensim. Лемматизация проводится при помощи пакета Spacy. Непосредственно модель также можно обучать при помощи библиотеки Genism.corpora.

Основой алгоритма определения тональности текста (рис. 3) в рекомендательной системе является применение модели на основе нейронной сети [6, 7, 8]. К традиционной схеме применения нейронной сети были добавлены шаги по предварительной обработке текста. Это позволило улучшить общее качество модели.

Алгоритм определения тональности текста включает следующие шаги:

1. Предварительная обработка текста: удаление знаков препинания, приведение к нижнему регистру.
2. Для расчета векторных представлений слов проводится обучение и применение модели Word2Vec [9]. Модель Word2Vec предназначена для получения слов на естественном языке, может находить взаимосвязи между словами на основе предположения о том, что в похожих контекстах встречаются семантически близкие слова.
3. Отображение текста в массив токенов.
4. Обучение и применение модели на основе нейронной сети.

В разработанной рекомендательной системе обучение происходило на размеченных данных Twitter. Записи содержались в 2 файлах — позитивные и негативные. Для токенизации и обучения нейронной сети использовалась библиотека Keras. Для исключения возможности переобучения использовался метод Dropout. В качестве функции ошибки использовалась бинарная кросс-энтропия:

$$H(P,Q) = -\sum_x P(x)\log Q(x).$$

Описание и результаты обучения и тестирования разработанной рекомендательной системы

Структурная схема разработанной рекомендательной системы соответствовала рис. 1. Распознавание тематики и тональности текста осуществлялось в соответствии с описанными алгоритмами. Для распознавания объекта на изображении использовалась предобученная модель Inception-v3 [10]. Это сверточная нейронная сеть, предназначенная для помощи в анализе изображений и обнаружения объектов. В качестве корректно определенных объектов была установлена граница в значение коэффициента равное 0.5. Всего из 1000 клас-

```
[(0,
  [('метро', 0.08333333333333333),
   ('понять', 0.08333333333333333),
   ('матрица', 0.08333333333333333),
   ('бета', 0.08333333333333333),
   ('станция', 0.08333333333333333),
   ('прикол', 0.08333333333333333),
   ('ладожский', 0.08333333333333333),
   ('воспроизвести', 0.08333333333333333),
   ('краш', 0.08333333333333333),
   ('тест', 0.08333333333333333)])]
```

Рис. 4. Результат работы LDA модели

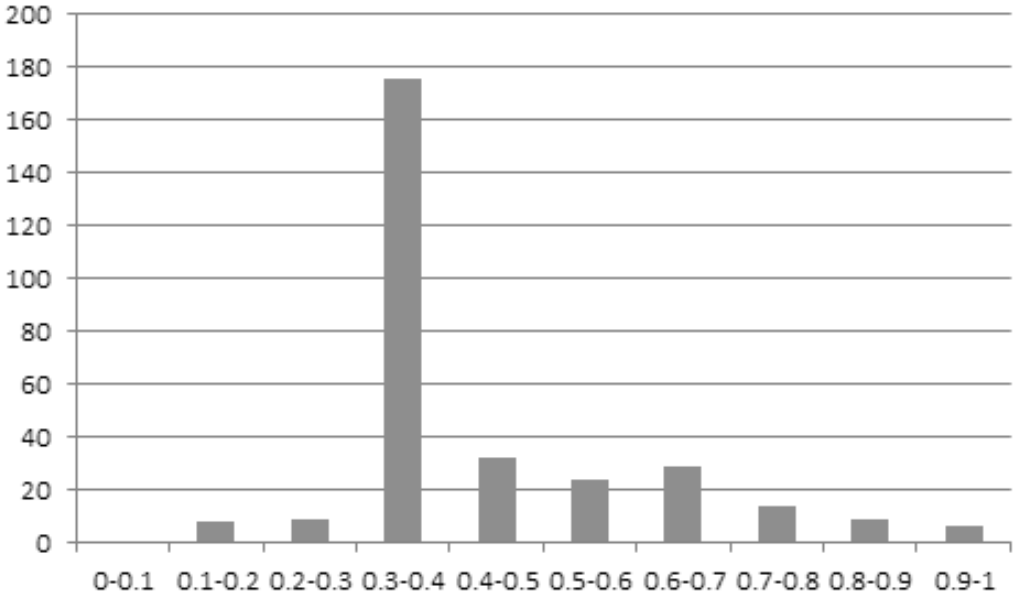

Рис. 5. Результаты работы модели для определения тональности текста

select photo.userid from photo_desc, photo, all_comment, comment where photo_desc.photo_desc like '%cat%' and photo_desc.photo_id=photo.id and comment.userid = all_comment.user_id and comment.comment = all_comment.comment and comment.photoid = photo.id and photo.userid in (select comment_main_theme.user_id from comment_main_theme where comment_main_theme like '%кот%') and photo.userid in (select user.id from user, photo where user.id = photo.userid and photo.longitude > 30.432 and photo.longitude < 30.434 and photo.latitude > 60.063 and photo.latitude < 60.065) and photo.userid in (select userid from rating where photoid in (select photo.id from photo_desc, photo where photo_desc.photo_desc like '%cat%' and photo_desc.photo_id=photo.id) and islike = 1) and tone > 0.3 and score > '0.5';

Рис. 6

сов модели на 127 изображениях было распознано 87 образов.

Результат работы LDA модели для определения основной тематики текста оценивался при помощи метрики согласованности. Значение метрики составило 0.659242705. Пример работы модели представлен на рис. 4. Видим основные тематики текстовых записей, которые принадлежат одному пользователю. Значение коэффициента показывает соотношение тематик в его записях.

Обучение модели выполнялось в течение 10 эпох. В качестве меры использовалась F-мера, которая учитывает значения Precision и Recall:

$$F = 2\frac{Precision * Recall}{Precision + Recall}$$

Результаты работы модели для определения тональности текста представлены на рис. 5. На оси абсцисс отображен процент, предсказанный моделью, а по оси ординат — количество подобных комментариев. Таким образом, большая часть всего набора текста носит нейтральный характер.

Все результаты работы моделей были записаны в базу данных. Также была записана информация о геотегах и рейтингах без дополнительной предобработки. Рассмотрим возможность работы с такой базой данных для получения информации о целевой аудитории. Предположим, что необходимо выявить целевую аудиторию владельцев котов для рекламы зоомагазина, который открылся в определенной точке. Чтобы использовать всю доступную информацию и получить гарантированных владельцев котов будем искать людей, которые публикуют фотографии с котами, отмечают фотографии с ними и положительно/нейтрально комментируют такие записи, а также упоминают их в своих комментариях. Для выполнения всех поставленных условий использовался следующий SQL-запрос (рис. 6).

Разработанная система на данный момент корректно функционирует на выгруженных обезличенных данных из реального проекта. Происходит подготовка к интеграции системы в проект.

Заключение

В результате разработанная система позволяет определять интересы пользователей на основе заданных параметров. Разработанная архитектура параллельной обработки различных наборов данных и последующее объединение в базе данных по уникальному идентификатору позволяет легко добавлять новые модули. Для возможности работы с небольшими наборами данных были использованы предобученные модели и размеченные источники, которые доступны в свободном доступе и могут использоваться в небольших проектах. Хранение результатов обработки в базе данных позволяет получать практически любые целевые аудитории на основе простого SQL-запроса.

ЛИТЕРАТУРА

1. Covington P., Adams J., Sargin E. Deep Neural Networks for YouTube Recommendations. Proceedings of the 10th ACM Conference on Recommender Systems, ACM, New York, NY, USA (2016).
2. Chandrashekar A., Amat F., Basilico J, Jebara T. Artwork Personalization at Netflix. The Netflix Tech Blog, 2017.
3. Ermakov, N.V., Molodyakov, S. A. Development and implementation of accelerated methods of data access. Journal of Physics: Conference Series 1326(1),012025 2019
4. Blei D., Ng Y., Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (2003) 993–1022.
5. Коляда А.С., Яковенко В. А., Гогунский В. Д.: Применение латентного размещения Дирихле для анализа публикаций из наукометрических баз данных. Труды Одесского политехнического университета. 2014. № 1. С. 186–191.
6. Subramanian V.: Deep Learning with PyTorch A practical approach to building neural network models using PyTorch / V. Subramanian — Packt Publishing, 2018 ISBN1788624335.
7. Sapozhnikova L.E., Gordeeva O. A.: Text classification using convolutional neural network. CEUR Workshop Proceedings. 2019. Vol. 2416. P. 219–226.
8. Ptukhin A.A., Khrushkov A. E., Bozhko E. M. Machine learning in the processing and analysis of texts. Сборник «Язык в сфере профессиональной коммуникации», 2019. С. 517–523.
9. Мишенин А.Н., Нефедова Е. А. Анализ тональности текстов с использованием технологии WORD2VEC. Естественные и математические науки в современном мире. 2016. № 7 (42). С. 89–97.
10. Joshi K., Tripathi V., Bose C., Bhardwaj C. Robust Sports Image Classification Using InceptionV3 and Neural Networks. Procedia Computer Science. International Conference on Computational Intelligence and Data Science, ICCIDS2019. 2020. С. 2374–2381.

© Монастырев Виталий Викторович (vit34-95@mail.ru), Молодяков Сергей Александрович (molodyakov_sa@spbstu.ru).

Журнал «Современная наука: актуальные проблемы теории и практики»