3. Tang X. et al. Li-ion battery parameter estimation for state of charge //American Control Conference (ACC), 2011. – IEEE, 2011. – С. 941-946.

А.В. Серебряков, Л.С. Зеленко, Д.С. Оплачко

ИССЛЕДОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ И РАЗРАБОТКА СИСТЕМЫ АВТОМАТИЧЕСКОГО РУБРИЦИРОВАНИЯ ТЕКСТОВ

(Самарский университет)

Обработка естественного языка (NLP, Natural language processing) – одно из направлений машинного обучения, целью которого является обработка и анализ больших массивов текстов на естественном языке. Это направление актуально, так как Интернет содержит огромное количество различных текстовых данных, с которыми человек уже давно перестал самостоятельно справляться.

Рубрицирование текстов – отнесение текста к одной из заранее известных тематических рубрик [1]. Такая задача решается, например, при отнесении новости или статьи к определенной рубрике (тематике) на информационных порталах, или для разделения заявок в системах технической поддержки по определенным проблемам.

Объемы неструктурированной текстовой информации, как и количество интернет-пользователей постоянно растут, следовательно, растет и актуальность разработки автоматизированного решения задачи рубрицирования, которое разделяло бы большие наборы текстов по определенным рубрикам, таким образом структурируя их и упрощая поиск нужной информации.

Выделяются два основных класса методов автоматического рубрицирования текстов – инженерные методы (методы, основанные на знаниях) и методы, основанные на машинном обучении. При применении инженерных методов, массив текстов разбивается по рубрикам с помощью формальных описаний каждой из рубрик, которые создаются лингвистами и экспертами в различных предметных областях. При применении методов, основанных на машинном обучении, производится статистический анализ коллекции документов, предварительно распределенных по рубрикам вручную, на основании которого образы рубрик строятся автоматически [2]. Под образами рубрик понимаются их формальные описания.

Правила рубрицирования текстов обычно основываются на наличии или отсутствии в текстах тех или иных лексических единиц. В простейшем случае правило отнесения текста к рубрике представляет собой дизъюнкцию наличия в тексте некоторых слов. В более сложном случае используются конъюнкции (требуется одновременное наличие двух или более слов) и отрицание (требуется отсутствие в тексте определенных слов) [2].

Задача рубрицирования является подвидом задачи классификации, так как относится только к массивам текстов и связана с разбиением их на рубрики

(тематики), в то время как под задачей классификации понимается отнесение *любых* объектов к *любым* заранее заданным категориям.

Преимущество методов решения задачи рубрицирования текста с помощью машинного обучения заключается в том, что они позволяют понизить трудоемкость разработки, так как не требуют работы лингвистов и экспертов в различных предметных областях над составлением образов рубрик [2]. При этом сложность использования таких методов заключается в том, что они требуют составления большой коллекции обучающих выборок, на основе которых алгоритм строит образы рубрик и «учится» относить текст к той или иной рубрике.

Рассмотрим формальную постановку задачи рубрицирования. Пусть X – множество текстовых документов, Y – конечное множество рубрик. Существует неизвестная целевая зависимость – отображение $y^*: X \to Y$, значения которой известны только на объектах конечной обучающей выборки $X^m = \{(x_1, y_2),..., (x_m, y_m)\}$. Требуется восстановить зависимость, т.е. построить функцию (алгоритм) $f: X \to Y$, предсказывающую $y \in Y$: y = f(x), то есть позволяющую отнести произвольный объект $x \in X$ к той или иной рубрике $y \in Y$.

Документы из обучающей выборки разделяются на два независимых набора: обучающего и тестового. С помощью обучающего набора алгоритм строит образы рубрик, анализируя входные данные и ответы к ним, и выявляя общие закономерности, присущие текстам на естественном языке. Тестовый набор данных нужен для оценки точности построенных образов. Каждый документ из тестового набора подается на вход построенному алгоритму $f: X \to Y$, а затем результат сравнивается с соответствующим ответом из тестовой выборки. Оценка качества при этом определяется как процент ошибки ответов на всей тестовой выборке.

Для разработки системы автоматического рубрицирования текстов и исследования были выбраны метод опорных векторов и многослойный персептрон, так как они считаются наиболее эффективными в задачах классификации текстов.

В качестве архитектуры для разрабатываемой системы была выбрана трехзвенная клиент-серверная архитектура «клиент – сервер приложений – БД». На сервере приложений будет развернута подсистема обучения, подсистема загрузки и предобработки обучающей выборки, подсистема работы с обученными моделями. Сервер приложений будет реализован с помощью языка Python, библиотек NLTK и Gensim. Интерфейс пользователя будет реализован в виде «тонкого» клиента на языке Typescript с использованием фреймворка Angular 11. Клиент с помощью веб-браузера будет взаимодействовать с сервером приложений. В БД будут храниться параметры для сохраненных обученных моделей.

На рисунке 1 приведена диаграмма вариантов использования системы со стороны пользователя.

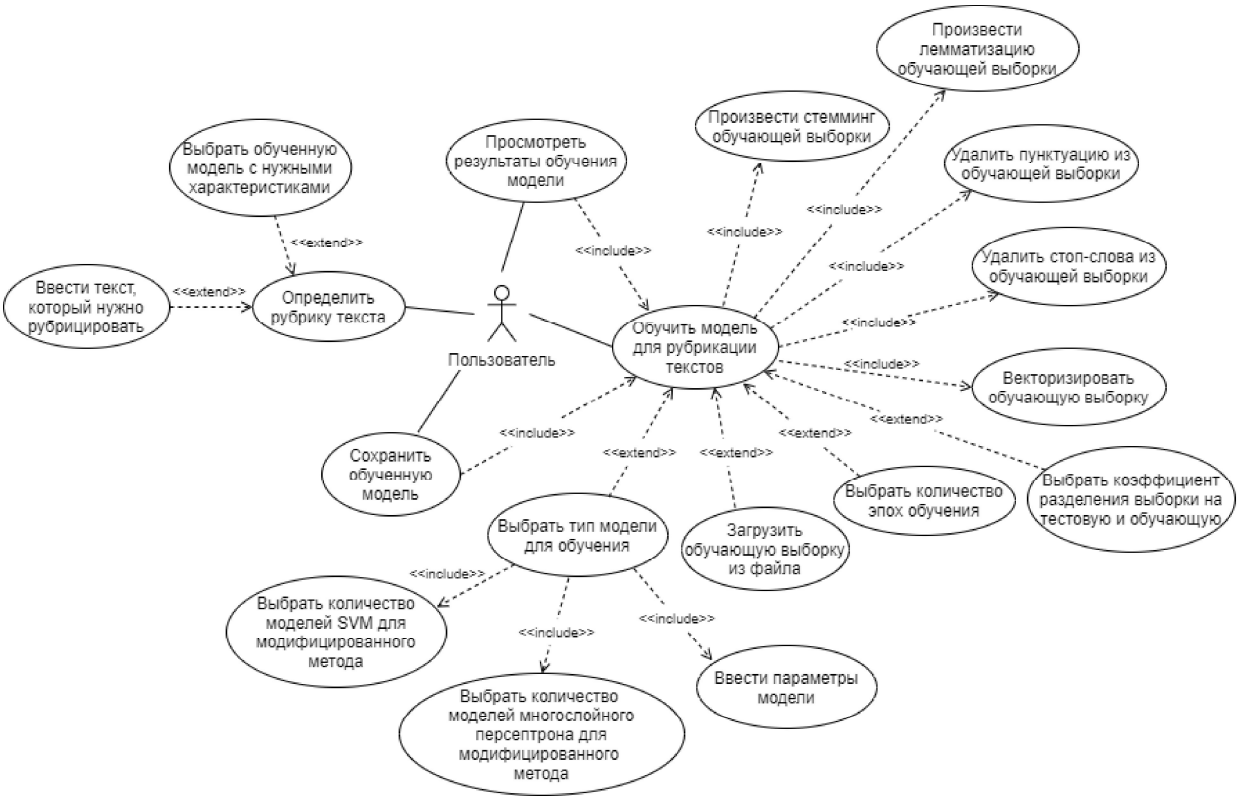

Рисунок 1 – Диаграмма вариантов использования для пользователя системы

Пользователь сможет обучить новую модель или определить рубрику для введенного текста с помощью уже обученных моделей, которые будут храниться на сервере приложений.

При обучении новой модели пользователь должен будет выбрать тип модели (SVM, многослойный персептрон или модифицированный), загрузить из файла обучающую выборку, в которой каждому набору текста сопоставлена рубрика, выбрать количество эпох обучения и коэффициент разделения выборки на тестовую и обучающую. Также пользователь сможет произвести стемминг, лемматизацию, векторизацию, удаление стоп-слов и пунктуации из обучающей выборки. После обучения модели пользователь сможет просмотреть результаты обучения и сохранить их.

Кроме метода опорных векторов и многослойного персептрона, в системе будет реализован гибридный алгоритм по принципу ансамбля методов. Пользователь при обучении сможет выбрать количество моделей SVM и количество моделей многослойного персептрона, таким образом получая набор из моделей разных методов, определяющих рубрику по принципу голосования.

Кроме разработки системы автоматической рубрикации текстов в рамках работы будет произведено исследование методов SVM и многослойного персептрона. Планируется исследовать, какие комбинации предобработки обучающей выборки, на каких параметрах для каждого из методов машинного обучения и какие комбинации моделей для модифицированного алгоритма дадут наибольшую точность обучения.

На текущий момент система находится на этапе реализации.

Литература

1 Автоматическая обработка текстов на естественном языке и анализ данных / Е.И. Большакова, К.В. Воронцов, Н.Э. Ефремова, Э.С. Клышинский. М.: НИУ ВШЭ, 2017. 269 с.

2 Добров А.В. Автоматическая рубрикация текстов средствами комплексного лингвистического анализа [Электронный ресурс]. http://www.aiire.org/pubs.php (дата обращения: 23.03.2021).

Д.А. Сорокин

АВТОМАТИЗАЦИЯ РАЗМЕЩЕНИЯ ДАТЧИКОВ УМНОГО ДОМА

(Самарский университет)

На сегодняшний день, благодаря стремлению человека к комфорту, явно можно видеть тренд к автоматизации жилых и производственных помещений – так называемые умные дома. Такие помещения и даже дворовые территории автоматически включают и выключают свет, следят за температурой, влажностью и обеспечивают безопасность. Для реализации этих функций используются различные сенсорные устройства (СУ), которые считывают информацию из окружающей среды. На основе данных этих устройств система выполняет те или иные действия. Для того чтобы СУ передавали достоверные данные важны их правильное размещение и, для некоторых типов датчиков, максимизация покрытия сканируемого пространства.

При наличии одинаковых сенсорных устройств (СУ) задача их размещения сводится к очевидному детерминированному решению, типа задачи упаковки шаров. Однако будем считать, что сенсорные устройства неоднородны, то есть имеют разный радиус действия, форму покрытия и функционально предназначены для сбора разной информации.

Задачу оптимального размещения сенсорных устройств в пространстве помещения можно свести к задаче, геометрического покрытия которая является частным случаем задачи оптимального проектирования и принадлежит к классу задач «раскроя и упаковки». Требуется расположить различные СУ, каждый из которых имеет свою зону покрытия, которая может отличаться как размером, так и формой, на покрываемой поверхности таким образом, чтобы вся или указанная часть поверхности была покрыта целиком. При этом в покрываемом пространстве могут находится препятствия, которые ограничивают зону покрытия СУ. Критериями оптимальности такой задачи является наименьшая площадь перекрытий зон покрытия СУ, использование минимального количества СУ и максимально возможное покрытие заданной области зонами покрытия СУ.

На сегодняшний день существуют исследования алгоритмов позволяющих размещать некоторые геометрические фигуры с максимизацией покрытия