1

МОДЕЛИРОВАНИЕ И АНАЛИЗ ДАННЫХ

НАУЧНЫЙ ЖУРНАЛ

MODELLING AND DATA ANALYSIS

SCIENTIFIC JOURNAL

2022

ISSN: 2219-3758
ISSN: 2311-9454 (ONLINE)

МОДЕЛИРОВАНИЕ И АНАЛИЗ ДАННЫХ

НАУЧНЫЙ ЖУРНАЛ

2022 • Том. 12 • № 1

MODELLING AND DATA ANALYSIS

SCIENTIFIC JOURNAL

2022 • Vol. 12 • № 1

Московский государственный
психолого-педагогический университет
Moscow State University
of Psychology & Education

Издаётся с 2011 года

Учредитель
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Московский государственный психолого-педагогический университет»

Адрес редколлегии:
г. Москва, ул. Сретенка, 29, факультет информационных технологий
Тел.: +7 (499) 167-66-74
E-mail: mad.mgppu@gmail.com

Журнал зарегистрирован в Государственном комитете РФ по печати.
Свидетельство о регистрации средств массовой информации
ПИ № ФС77-52058 от 7 декабря 2012 года

ISSN: 2219-3758
ISSN: 2311-9454 (online)

СОДЕРЖАНИЕ

АНАЛИЗ ДАННЫХ

МЕТОДЫ ОПТИМИЗАЦИИ

МЕТОДИКА ОБУЧЕНИЯ

CONTENTS

Моделирование и анализ данных
2022. Том 12. № 1. С. 5–15
DOI: https://doi.org/10.17759/mda.2022120101
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2022. Vol. 12, no. 1, pp. 5–15
DOI: https://doi.org/10.17759/mda.2022120101
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

АНАЛИЗ ДАННЫХ

УДК 004.855

Кластеризация водителей по степени опасности вождения с использованием алгоритмов машинного обучения

Баданина Н.Д.*
Московский авиационный институт (МАИ)
г. Москва, Российская Федерация
ИПМ им. М.В.Келдыша РАН
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-5301-1526
e-mail: natashabadanina99@gmail.com

Судаков В.А.**
Московский авиационный институт (МАИ)
г. Москва, Российская Федерация
ИПМ им. М.В.Келдыша РАН
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-1658-1941
e-mail: sudakov@ws-dss.com

Для цитаты:
Баданина Н.Д., Судаков В.А. Кластеризация водителей по степени опасности вождения с использованием алгоритмов машинного обучения // Моделирование и анализ данных. 2022. Том 12. № 1. С. 5–15. DOI: https://doi.org/10.17759/mda.2022120101

***Баданина Наталья Дмитриевна,** студент магистратуры, Московский авиационный институт (национальный исследовательский университет) (МАИ (НИУ)), программист, Федеральное государственное учреждение «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук» (ИПМ им. М.В.Келдыша РАН), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-5301-1526, e-mail: natashabadanina99@gmail.com

****Судаков Владимир Анатольевич,** доктор технических наук, профессор, Московский авиационный институт (национальный исследовательский университет) (МАИ (НИУ)), ведущий научный сотрудник, Федеральное государственное учреждение «Федеральный исследовательский центр Институт прикладной математики им. М.В. Келдыша Российской академии наук» (ИПМ им. М.В.Келдыша РАН), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-1658-1941, e-mail: sudakov@ws-dss.com

В работе проведено исследование по определению опасности вождения транспортного средства посредством анализа сигналов, полученных во время поездки. Был применен функционал ряда современных моделей кластеризации для разбиения множества водителей на классы, соответствующие степени опасности вождения. Разработан новый подход агрегации данных с целью кластеризации с использованием гистограмм распределения сигналов. Полученные результаты могут быть использованы в промышленных системах мониторинга качества поведения водителей в ретроспективе.

Ключевые слова: кластеризация, гистограммы распределения, опасное вождение, системы мониторинга, машинное обучение.

1. ВВЕДЕНИЕ

Системы мониторинга безопасности вождения транспортных средств (ТС) призваны снижать аварийность на дорогах; контролировать общее соблюдение правил дорожного движения. Частично разработки нацелены на заблаговременное предотвращение аварийных ситуаций. Такие системы безусловно важны, по причине того, что от них в прямом смысле может зависеть безопасность на дороге, жизни водителей, пассажиров и пешеходов. При должном контроле поведения водителя, существует возможность применения санкционных мер, с целью улучшить его поведение за рулем, что в перспективе может снизить риск дорожно-транспортного происшествия.

Методы анализа данных широко применяются для обработки больших данных высокой размерности. Кластеризация – разбиение множества объектов на группы (кластеры), основываясь на свойствах этих объектов. Кластер представляет собой группу объектов, имеющих общие признаки. Целью алгоритмов кластеризации является создание классов, которые максимально связаны внутри себя, но различны друг от друга [1]. Таким образом, в системах мониторинга опасности вождения можно применить алгоритмы кластеризации, с целью решения задачи группировки водителей на некоторое количество классов, соответствующее различным типам вождения.

2. РЕЛЕВАНТНЫЕ ИССЛЕДОВАНИЯ И РАЗРАБОТКИ

На рынке присутствуют как государственные реализации системы мониторинга, так и частные. К государственным относится система камер фото- и видео-фиксации ГИБДД. В их основе лежит одна из технологий искусственного интеллекта – компьютерное зрение. Эта система имеет ряд преимуществ, однако, не имеет возможности сделать вывод о безопасности вождения в любом моменте времени – необходимо обязательное наличие камер или иных средств фиксации нарушения. У органов регулировки дорожно-транспортной ситуации нет возможности обеспечить стопроцентное покрытие всех автомагистралей, дорог специализированным оборудованием, и нет возможности осуществлять мониторинг ТС на протяжении всего пути. Во многом операционная система Яндекс.Навигатор и подобные «помогают» водителям избежать нарушений, предупреждая о наличии средств контроля заранее. Это ведет

к недостоверной статистике о поведении водителя. Третьим недостатком является зависимость от технологии компьютерного зрения, которая не способна работать при ненадлежащем качестве изображения, при поломке фото- или видеорегистраторов.

Альтернативные способы контроля качества вождения представлены частными компаниями. Яндекс.Про считывает данные акселерометра устройства, на которое он установлен, – смартфона или планшета. Если водитель совершает агрессивные манёвры – резкий старт и торможение, многократную смену полос, несоблюдение дистанции, – акселерометр фиксирует перегрузку и считает нарушением при систематичности [3]. Недостатком является факт того, что по каждому ТС показатели собираются множество раз за все время поездки, что усложняет анализ и может вести к ложным выводам из-за ошибки оборудования или из-за факта разовых нарушений.

В исследовательской области есть ряд работ по определению опасного вождения – кластеризация водителей по стилям торможения [2] с использованием нейронной сети Кохонена; определение поведения водителей на основе мониторинга движения в реальном времени [5]; детекция несосредоточенности водителя с использованием изображения машины внутри и снаружи [6].

Таким образом, на рынке мониторинга безопасности вождения транспортных средств присутствует возможность внедрения новых технологий, использующих машинное обучение. В данной работе описана методика построения моделей кластеризации водителей для задачи обнаружения паттернов опасного вождения.

3. ГИСТОГРАММЫ РАСПРЕДЕЛЕНИЯ КАК СПОСОБ СОСТАВЛЕНИЯ ОБУЧАЮЩЕГО МНОЖЕСТВА

В качестве данных в работе был использован датасет, собранный из информации о сигналах с ТС. К сигналам относятся значения скорости; значения ускорений по трем осям x, y, z.

Дано множество водителей $D = \{d_1, d_2, \ldots, d_n\}$. Для каждого $d_i \in D$, $i = 1, \ldots, n$ дан вектор скоростей $V = (v_1, v_2, \ldots, v_m, \ldots)$, где каждое значение соответствует скорости ТС. Аналогично заданы вектора ускорений $A_x = (a_{x_1}, a_{x_2}, \ldots, a_{x_m}, \ldots)$, $A_y = (a_{y_1}, a_{y_2}, \ldots, a_{y_m}, \ldots)$, $A_z = (a_{z_1}, a_{z_2}, \ldots, a_{z_m}, \ldots)$ по осям x, y, z соответственно. Требуется разделить множество D на K непересекающихся кластеров $C = \{c_1, c_2, \ldots, c_k\}$, $k = 1, \ldots, K$.

Данные о сигналах поступали с некоторой периодичностью на протяжении всей поездки на ТС, даже при условии того, что фактическая скорость могла быть нулевой, то есть факт движения отсутствовал. На их основе были обучены и протестированы модели кластеризации, реализованные с помощью методов машинного обучения без учителя, то есть в данных отсутствовала разметка о принадлежности того или иного водителя к некоему классу, отвечающему за степень опасности вождения.

Данные по трекам были агрегированы. Для каждого водителя была собрана вся доступная информация по сигналам, записанным во время всех поездок. За каждую

поездку могло поступить множество сигналов с разным временным промежутком. Проблема обучения моделей на таких данных состоит в том, что для каждого водителя количество собранных сигналов велико и для разных водителей их может быть разное количество, и, следовательно, вектор признаков будет иметь различную длину. С векторами различной длины нельзя составить обучающее множество. Вариант дополнения каждого вектора до максимальной длины нецелесообразен, так как может вести к существенному увеличению времени обучения модели и ложным выводам.

Решением описанной проблемы является составление частотных гистограмм распределения значений сигналов по интервалам. Гистограммы распределения указывают насколько часто встречаются те или иные значения. Рассмотрим алгоритм построения для задачи кластеризации водителей.

Выделяются сигналы одной природы. В рассматриваемой задаче описано множество сигналов $S = \{V,\ A_x,\ A_y,\ A_z\}$. Для каждого вектора из множества S выделяются интервалы значений. Не допускается перекрытие промежутков и наличие пропущенных значений.

- Интервалы задаются экспертом;
- Весь диапазон значений разбивается на равное количество частей с некоторым шагом;
- Оптимальное количество интервалом определяется математически, исходя из мощности выборки. Применяется формула Стерджесса,

$$m = 1 + 3,322 * \lg(n) \tag{1}$$

где n – количество наблюдений. После подсчета коэффициента m весь диапазон значений от минимального до максимального разбивается на равные части. Ширина интервала определяется по формуле

$$w = \frac{X_{max} - X_{min}}{m} \tag{2}$$

Тогда первый интервал начинается в X_{min}, а последний (с номером m) заканчивается в X_{max}. Интервалы составляются следующим образом:

$$(X_{min},\ X_{min} + w), \ldots, (X_{max} - w,\ X_{\max}).$$

После определения интервалов осуществляется подсчет частоты попадания значений сигналов в соответствующие интервалы. Частота считается явно: сколько из всех значений находятся внутри интервала. Пусть a_i и b_i – левая и правая граница интервала i соответственно, тогда считается количество значений x_j, для которых выполняются неравенства $x_j \geq a_i$ и $x_j < b_i$.

Таким образом, можно подсчитать частоту попадания значений сигналов в определенный интервал и составить обучающее множество. После выполнения вышеописанного алгоритма был составлен датасет следующего вида: колонки соответствуют интервалу сигнала из S, строки – водителям из D, а цифры – количеству вхождений значений сигнала в интервал.

4441	634	442	209	251	231	310	326	263	42	...	21
21786	2642	4962	3853	3861	4174	3764	2443	1024	181	...	0
4565	1196	923	777	1362	1391	1908	1814	2254	466	...	0
4846	731	681	381	377	503	477	503	168	61	...	0
78748	17543	10267	3995	3309	3303	2567	919	86	1	...	0

Рис. 1. Пример полученного датасета

Интервалы были заданы одинаковой длины: для скорости с шагом 10, а для ускорений с шагом 100. Каждому водителю соответствует единственная строка, в которой указано количество попадания значений сигнала в заданные промежутки значений. Такой подход позволяет:

- учесть данные по всем поездкам, совершенным водителем ТС;
- создать вектор определенной неизменной длины для каждого водителя;
- объединять интервалы при необходимости сбора более обобщенной статистики;
- добавлять или исключать водителей без необходимости пересчета всего обучающего множества;
- собрать наглядную статистику по каждому доступному водителю за весь период активности;
- исключить введения штрафных санкций за разовые нарушения, которые не свойственны определенному водителю;
- перевести целочисленные значения частотности в долевые значения, указывающие какой процент значений попадает в определенный интервал по отношению ко всему множеству значений сигналов, которое было разбито на промежутки.

4. ПОСТРОЕНИЕ МОДЕЛИ КЛАСТЕРИЗАЦИИ

Цель моделей кластеризации – определить K количество групп для n объектов на основании метрик сходства таким образом, чтобы максимизировать схожесть между объектами одного класса, одновременно минимизировав схожесть между объектами разных классов.

Пусть $X = \{x_i\}$, $i = 1, \dots n$ – множество из n точек некоторой размерности d, которое необходимо разбить на K кластеров $C = \{c_k\}$, $k = 1, \dots, K$.

Алгоритм K-means стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров. Пусть μ_k – центр кластера c_k. Среднеквдаратическая ошибка между μ_k и точками кластера c_k определяется как

$$J(c_k) = \sum_{x_i \in c_k} x_i - \mu_k^2 \tag{3}$$

Цель модели K-means – минимизировать суммарную среднеквадратическую ошибку для всего множества кластеров K [4].

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} x_i - \mu_k^2 \rightarrow min \tag{4}$$

В работе были протестированы различные методы кластеризации, для каждого из которых была вычислена метрика силуэт. Коэффициент силуэт вычисляется с помощью среднего внутрикластерного расстояния и среднего расстояния до ближайшего кластера по каждому наблюдению. Силуэт вычисляется по формуле

$$s = \frac{b - a}{\max(a, b)} \tag{5}$$

где a – среднее расстояние от данного объекта до объектов из того же кластера, b – среднее расстояние от данного объекта до объектов из ближайшего кластера (отличного от того, в котором лежит сам объект). Эта метрика позволяет оценивать качество работы моделей кластеризации, обученных без учителя.

Таблица 1

Вариации данных

	Интервалы составлены с шагом 10 для скорости и с шагом 100 для ускорения
Частотные долевые значения в %	А
Частотные целочисленные значения	Б

Таблица 2

Результаты работы алгоритмов

Название модели с указанием основных параметров	**Метрика силуэт**	
	А	**Б**
KMeans n_clusters=5	0.459	0.505
KMeans А, Б: n_clusters=2, init='k-means++', max_iter=300, algorithm= 'auto'	0.686	0.801
AgglomerativeClustering n_clusters=5	0.449	0.434
AgglomerativeClustering А: n_clusters=2, affinity='l1', linkage='complete' Б: n_clusters=2, affinity='euclidean', linkage=' single'	0.682	0.892
AffinityPropagation damping=0.9	0.221	0.344
AffinityPropagation А: damping=0.8, max_iter=200, affinity='euclidean' Б: damping=0.85, max_iter=200, affinity='euclidean'	0.224	0.342

Название модели с указанием основных параметров	**Метрика силуэт**	
	А	**Б**
Birch threshold=0.01, n_clusters=5	0.446	0.434
Birch А: n_clusters=2, threshold=0.36, branching_factor=50 Б: n_clusters=2, threshold=0.1, branching_factor=50	0.686	0.804
DBSCAN А: eps=0.30, min_samples=9 Б: eps=5, min_samples=4	0.384	-0.281
DBSCAN А: eps=0.5, min_samples=8, algorithm='auto' Б: eps=9, min_samples=3, algorithm='auto'	0.473	-0.277
MiniBatchKMeans n_clusters=5	0.442	0.497
MiniBatchKMeans А: n_clusters=2, init='k-means++', max_iter=100, batch_size=4, reassignment_ratio=0.01 Б: n_clusters=2, init='k-means++', max_iter=100, batch_size=64, reassignment_ratio=0.01	0.686	0.802
MeanShift()	0.537	0.394
MeanShift max_iter=300, bandwidth=0.8	0.683	0.002
OPTICS А: eps=0.8, min_samples=10 Б: eps=10, min_samples=9	-0.643	-0.624
OPTICS А, Б: eps=1, min_samples=2, metric= 'euclidean', cluster_method= 'xi', algorithm= 'auto'	-0.291	-0.131
SpectralClustering n_clusters=5	0.444	-0.322
SpectralClustering А: n_clusters=2, eigen_solver='arpack', gamma=0.2, affinity='rbf', assign_labels='kmeans' Б: n_clusters=2, eigen_solver='lobpcg', affinity='nearest_neighbors', assign_labels='discretize'	0.686	0.352
GaussianMixture n_components=2	0.376	0.099
GaussianMixture А: covariance_type='spherical', n_components=2, init_params='kmeans' Б: covariance_type='tied', n_components=2, init_params='kmeans'	0.682	0.805

Из Таблицы 2 видно, что метрику силуэт удалось улучшить для большинства моделей. От вида данных, подаваемых в модель зависят полученные результаты, поэтому при разработке стоит учитывать разнообразие существующих алгоритмов, уделить внимание их подбору.

Рассмотрим подробнее модель KMeans на данных типа Б, так как она показала одну из наилучших метрик, а также проста в понимании. Построим гистограммы для центров кластеров, на которые модель разделила данные. Центр можно считать средним значением по кластеру. Из Рис. 2 видно, что 3 класс можно считать классом, относящимся к опасному вождению из-за распределения значений в интервалах с высокими скоростями.

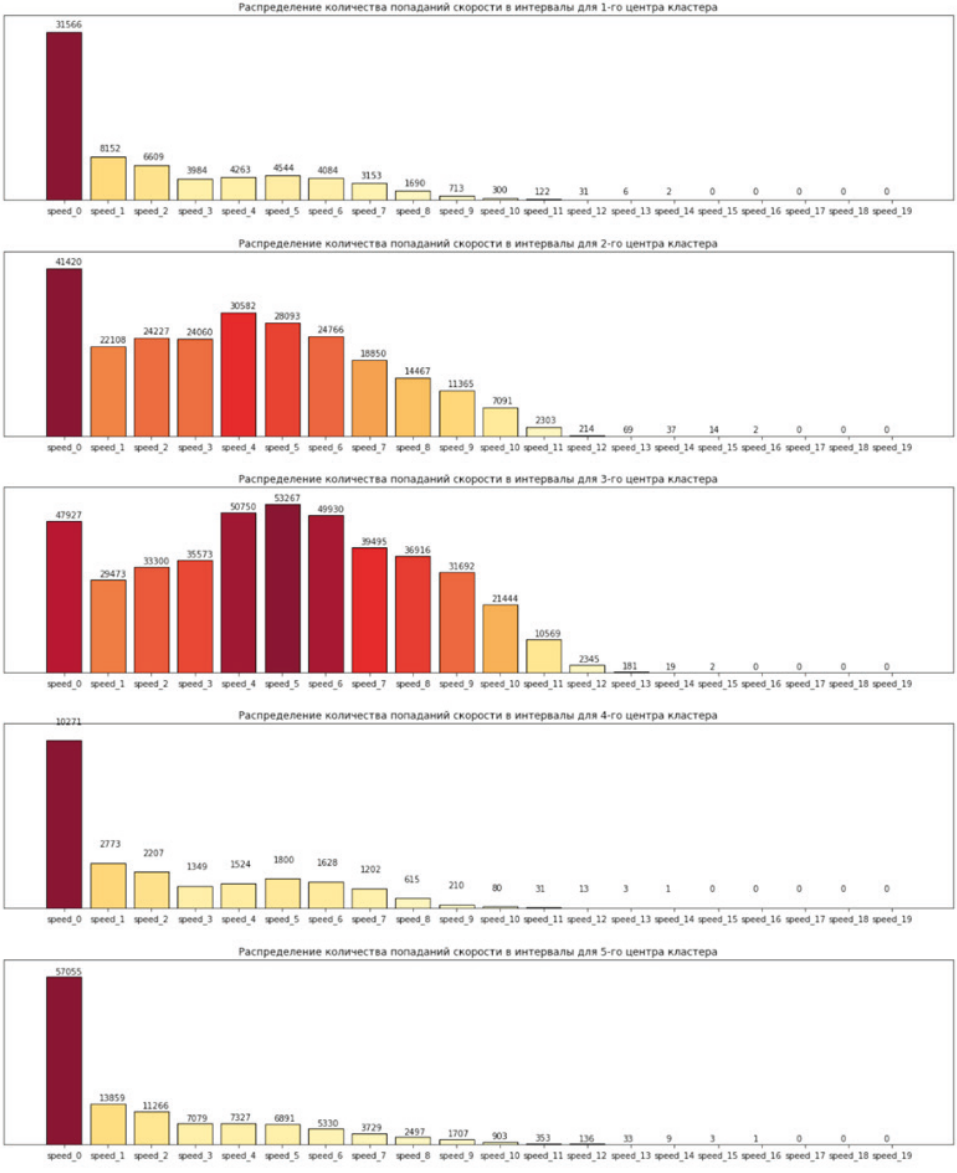

Рис. 2. Гистограммы для центров классов

Для выполнения расчетов и построения моделей был выбран язык программирования Python3, с использованием среды разработки Jupyter Notebook. Выбор обусловлен наличием большого количество библиотек с готовыми реализациями математического аппарата, а также наличием мощных инструментов для визуализации готовых результатов. В ходе выполнения работы были использованы готовые библиотеки, в том числе feather для хранения больших данных. Формат feather был использован как альтернатива csv, так как позволяет уменьшить объем сохраняемых файлов, увеличивает скорость чтения и записи данных. Были обучены модели кластеризации из широко используемого пакета Scikit Learn.

5. ЗАКЛЮЧЕНИЕ

Целью работы являлась реализация методов кластеризации в решении задачи разбиения множества водителей на классы, соответствующие степени опасности вождения. В качестве данных были использованы записи о сигналах водителей за некоторый временной период. Для агрегации датасета были использованы гистограммы распределения, позволяющие привести данные в формат, пригодный для обучения моделей машинного обучения, также, с помощью этого подхода была собрана собирательная статистика за все время активности водителя ТС.

В ходе выполнения работы был обучен ряд моделей кластеризации и подсчитала метрика силуэт. При принятии решения о качестве модели кластеризации, обученной без учителя, можно опираться на значение этой метрики в том случае, когда результаты работы модели совпадают с ожиданиями от поставленной задачи.

Использование описанного метода кластеризации и работы с гистограммами распределения сигналов в системах мониторинга качества вождения может улучшить контроль за поведением водителей, поможет выявлять и предотвращать поведение, способное привести к дорожно-транспортному происшествию.

Литература

1. *Кутуков Д.С.* Применение методов кластеризации для обработки новостного потока // Технические науки: проблемы и перспективы: материалы I Междунар. науч. конф. Санкт-Петербург: Реноме, 2011. С. 77–83.
2. *Дик Д.И.* Кластеризация водителей по стилям торможения // Курганский государственный университет. Вестник КГУ. 2012. № 2(24). С. 17–20.
3. Мониторинг манеры вождения [Электронный ресурс]: https://pro.yandex/ru-ru/moskva/knowledge-base/taxi/safety/monitoring-driving
4. *Jain A.K.* Data clustering: 50 years beyond K-means // Pattern Recognition Letters, 31(8). 2010. pp. 651–666.
5. *Huaikun Xiang , Jiafeng Zhu , Guoyuan Liang and Yingjun Shen.* Prediction of Dangerous Driving Behavior Based on Vehicle Motion State and Passenger Feeling Using Cloud Model and Elman Neural Network // Frontiers in Neurorobotics. April 2021. Vol. 15. p. 16.
6. *Omerustaoglu Furkan, Sakar C. Okan, Kar Gorkem.* Distracted driver detection by combining in-vehicle and image data using deep learning // Applied Soft Computing 96(6). 2020.
7. *J. Hartigan.* Clustering Algorithms // New York: Wiley, 1975.

Driver Clustering According to the Ratio of Dangerous Behavior Using Machine Learning Algorithms

Natalya D. Badanina*
Moscow Aviation Institute (MAI), Moscow, Russia
Keldysh Institute of Applied Mathematics, Moscow, Russia
ORCID: https://orcid.org/0000-0002-5301-1526
e-mail: natashabadanina99@gmail.com

Vladimir A. Sudakov**
Moscow Aviation Institute (MAI), Moscow, Russia
Keldysh Institute of Applied Mathematics, Moscow, Russia
ORCID: https://orcid.org/0000-0002-1658-1941
e-mail: sudakov@ws-dss.com

The paper conducts the research of defining dangerous driving of a vehicle using signals collected during the ride. A number of modern clustering models for drivers segmentation on classes based on the ratio of dangerous driving was used. New approach of data aggregation aiming to cluster data by signal distribution histograms was developed. Achieved results could be used in commercial systems that monitor the quality of drivers behavior in retrospective.

Keywords: clustering, distribution histograms, dangerous driving, monitoring systems, machine learning.

For citation:
Badanina N.D., Sudakov V.A. Driver Clustering According to the Ratio of Dangerous Behavior Using Machine Learning Algorithms. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*, 2022. Vol. 12, no. 1, pp. 5–15. DOI: https://doi.org/10.17759/mda.2022120101 (In Russ., abstr. in Engl.).

References

1. Kutukov D.S. Primenenie metodov klasterizacii dlya obrabotki novostnogo potoka // Tekhnicheskie nauki: problemy i perspektivy: materialy I Mezhdunar. nauch. konf. Sankt-Peterburg: Renome, 2011. pp. 77–83. (In Russ.).
2. Dik D.I. Klasterizaciya voditelej po stilyam tormozheniya. Kurganskij gosudarstvennyj universitet. Vestnik KGU. 2012. № 2(24). pp. 17–20. (In Russ.).
3. Monitoring manery vozhdeniya [URL]: https://pro.yandex/ru-ru/moskva/knowledge-base/taxi/safety/monitoring-driving. (In Russ.).

***Natalya D. Badanina,** Master Student, Moscow Aviation Institute (National Research University), Programmer, Keldysh Institute of Applied Mathematics (Russian Academy of Sciences), Moscow, Russia, ORCID: https://orcid.org/0000-0002-5301-1526, e-mail: natashabadanina99@gmail.com

****Vladimir A. Sudakov,** Doctor of Technical Sciences, Professor, Moscow Aviation Institute (National Research University), Leading Researcher, Keldysh Institute of Applied Mathematics (Russian Academy of Sciences), Moscow, Russia, ORCID: https://orcid.org/0000-0002-1658-1941, e-mail: sudakov@ws-dss.com

4. Jain A.K. Data clustering: 50 years beyond K-means // Pattern Recognition Letters, 31(8). 2010. pp. 651–666.
5. Huaikun Xiang, Jiafeng Zhu , Guoyuan Liang and Yingjun Shen. Prediction of Dangerous Driving Behavior Based on Vehicle Motion State and Passenger Feeling Using Cloud Model and Elman Neural Network // Frontiers in Neurorobotics. April 2021. Vol. 15. p. 16.
6. Omerustaoglu Furkan, Sakar C. Okan, Kar Gorkem. Distracted driver detection by combining in-vehicle and image data using deep learning // Applied Soft Computing 96(6). 2020.
7. J. Hartigan. Clustering Algorithms // New York: Wiley, 1975.

Получена 04.03.2022
Принята в печать 14.03.2022

Received 04.03.2022
Accepted 14.03.2022

Моделирование и анализ данных
2022. Том 12. № 1. С. 16–26
DOI: https://doi.org/10.17759/mda.2022120102
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2022. Vol. 12, no. 1, pp. 16–26
DOI: https://doi.org/10.17759/mda.2022120102
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

УДК 519.876.2

О разрешении проблемы оценивания вероятности выполнения разработанного плана

Воронов М.В.*
Московский государственный психолого-педагогический университет (ФГБОУ ВО МГППУ), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0001-7839-6250
e-mail: mivoronov@yandex.ru

Рассматриваются вопросы оценки вероятности реализации разработанного плана деятельности предприятия. Предлагается использование конструктивно-имитационного метода, обеспечивающего возможность на единой методической основе разработки программных комплексов автоматической выработки планов с полным набором оценок характеризующих их показателей, включая и вероятность реализации плана.

Ключевые слова: план, конструктивный процесс, модель, имитация, эффективность, вероятность.

Для цитаты:
Воронов М.В. О разрешении проблемы оценивания вероятности выполнения разработанного плана // Моделирование и анализ данных. 2022. Том 12 № 1. С. 16–26. DOI: https://doi.org/10.17759/mda.2022120102

1. ПОСТАНОВКА ЗАДАЧИ

Рациональная деятельность людей предусматривает предварительное обдумывание своих последующих действий с тем, чтобы достичь преследуемую цель. Результатом этого обдумывания, как правило, является алгоритм последующих действий, который называют планом, а собственно процесс его разработки–планированием.

Разработка плана всегда связана с формированием представлений о будущем (с «построением моста», который обеспечивает дорогу из настоящего в будущее) и призвана решать многочисленные вопросы анализа ситуации, целеполагания, формирования упорядоченной последовательности действий, их оценки, а также создания условий, способствующих осуществлению этих действий [1].

***Воронов Михаил Владимирович,** доктор технических наук, заведующий кафедрой прикладной математики факультета Информационных технологий, Московский государственный психолого-педагогический университет (ФГБОУ ВО МГППУ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0001-7839-6250, e-mail: mivoronov@yandex.ru

Планирование осуществляется в условиях целого спектра неопределенностей. В частности, это неопределенности в знаниях о сложившейся ситуации (они, как правило, не полны и не точны) и того, как она будет изменяться со временем [2]. В этих условиях и требуется проложить траекторию движения объекта рассмотрения в некотором (заданном или сконструированным) фазовом пространстве от заданного исходного состояния до целевого, устраняя неопределенность за счет введения дополнительной информации.

Как правило, эта траектория представляет собой упорядоченную последовательность этапов, и для формирования каждого из них необходимо выработать так называемое частное решение. При выработке очередного частного решения сталкиваются с проблемой формирования множества его потенциальных вариантов (какой шаг делать следующем?). Обычно мощность этого множества огромна (часто бесконечен), и полный их перебор практически невозможен. Для преодоления такого рода трудностей при разработке важных планов в качестве предварительного вводят этап, называемый выработкой замысла.

Важно заметить, что выработка замысла−глубоко творческий процесс и осуществляется он исключительно субъектом. Именно субъект, наделенный полномочиями, обладая определенными знаниями и опытом, а также информацией о сложившейся ситуации и закономерностях ее изменений, должен актуализировать подлежащую достижению цель, сформулировать способ и обобщенную схему ее достижения.

На практике замысел в явном виде не всегда выделяется в качестве первого этапа планирования. При выработке плана субъект (индивидуальный или коллективный) «вручную» объединяет решение всех задач в единый процесс, сужая при этом число подлежащих рассмотрению вариантов очередного шага. При отсутствии форс-мажорных обстоятельств, роль замысла исполняют предварительно утвержденные правила, рекомендации и схемы выработки частных решений в типовых ситуациях.

На промышленных предприятиях роль замысла играют так называемые «рабочие планы». Рассмотрим один весьма обобщенный пример [3]. Предприятие, получив заказ на изготовления некоторого заказа, вначале рассматривает возможность его выполнить за счет наличия готовой продукции. Для этого нужно определить, какое количество изделий в данной ситуации целесообразно взять непосредственно со склада готовой продукции. Дело в том, что полностью выполнять заказ за счет готовой продукции не всегда является лучшим решением. Поэтому здесь используют наработанные опытом правила, например, такое: «при исполнении большого заказа со склада готовой продукции данного вида допустимо брать не более 90 % от заказанного числа единиц продукции». Если принятое в рамках такого рода правил частное решение не закрывает вопрос планирования исполнения заказа, формируется решение о собственно изготовлении оставшейся части изделий (как и когда оставшуюся часть заказа можно в сложившихся условиях изготовить). Когда соответствующее решение сформулировано, встает вопрос об обеспеченности соответствующего производства необходимыми ресурсами. Если в сложившейся ситуации части потребных ресурсов недостаточно или не целесообразно их брать, то рассматривается вопрос о приоб-

ретении ресурсов. Тем самым возникают вопросы выделения финансовых средств, закупки и доставки ресурсов.

Выработка замысла, как результата интеллектуальной деятельности индивидуального или коллективного субъекта, является обязательным условием начала формирования собственно плана. Далее будем предполагать, что замысел выработан, т.е. известна цель и состояние исходной ситуации, а обобщенные правила и схемы принятия частных решений в конкретных условиях в достаточно полной мере описаны. Подчеркнем, что и в этом случае необходимо формирование плана, как достаточно детально и четко описанной упорядоченной последовательности шагов на пути достижения поставленной цели.

Поскольку вариантов планов достижения цели может быть построено, вообще говоря, более одного, то стремятся принять тот, который в смысле принятого критерия (критериев) является наилучшим. Естественно, в число критериев в первую очередь включаются так называемые результирующие показатели, т.е. показатели характеризующие целевой эффект. Это количество единиц продукции, получаемая прибыль, рентабельность, время выполнения заказа и др. Заманчиво разработать план лучший с позиций всех этих показателей. Поскольку принимаемые во внимание показатели, как правило, противоречивы, множество планов, для которых все частные критерии достигают экстремальных значений, обычно оказывается пустым (общего решения многокритериальной задачи не существует). Например, увеличение выпуска изделий связано, как правило, с увеличением затрат ресурсов, поэтому требовать выработки наибольшего количества изделий при минимизации затрат – абсурдно. При необходимости же решения многокритериальной задачи стремятся свести ее к задаче однокритериальной, сворачивая вектор критериев в скаляр или вводя дополнительную информацию, назначая, например, один критерий главным, а другие превращают в ограничения. Тем самым вносится дополнительная информация и решению подлежит иная задача.

Наряду с результирующими показателями существенно важно иметь оценки выполнимости разработанного плана, центральное место среди которых занимает вероятность выполнения плана, ибо какие бы планы не разрабатывались, они немногого стоят, если имеют мало шансов на свою реализацию.

В настоящее время имеется целый арсенал средств, позволяющих как-то оценивать выполнимость плана. Обычно это осуществляется путем прямых расчетов различного рода ресурсов или следуют советам экспертов, направленных на увеличение вероятности реализации отдельных составляющих плана, могут использоваться и инструменты вспомогательного назначения, например, формирования план-графиков выпуска деталей, план-графиков обеспечения материалами и комплектующими, сбыта, контроля всех этапов работы предприятия и т.п. [4, 5]. В то же время работ посвященным оценке вероятности разработанного плана крайне мало, что и обусловливает актуальность рассматриваемой темы.

Почему же оценка вероятности плана практически никогда не фигурирует при принятии решений? По-видимому, из-за сложности этой задачи. Во-первых, учет

вероятности исполнения плана наравне с целевыми показателями приводит к задаче векторной оптимизации, не имеющей, вообще говоря, общего решения. Во-вторых, в ходе разработки достаточно детальных планов необходимо рассматривать и принимать множество частых (текущих) решений, когда оценить количественное влияние каждого такого шага на конечный результат крайне проблематично. Существуют различные методы, основанные на вероятностно-статистических методах. Однако для их применения необходимо знать вероятностные характеристики составляющих и уметь строить на их основе аналитические модели оценки вероятности реализации плана, что для сложной деятельности весьма затруднительно или не обеспечивает получение результатов с необходимой точностью [6]. Остается метод статистических испытаний. Однако во многих случаях его прямое использование не дает ожидаемого эффекта. Дело в том, что, если разработанный план не носит характер конструктивного процесса, практически любое отклонение сымитированного исполнения компонента плана приводит к необходимости считать весь план неисполненным, а перепланирование недопустимо. Но такой подход противоречит существующему положению дел: на практике широко используется оперативное управление, ориентированное на принятие мер по возвращению процесса на запланированную траекторию.

2. КИМ-ПЛАНИРОВАНИЕ

Оценка вероятности исполнения плана может быть удовлетворительно решена при использовании метода конструктивно-имитационного моделирования (КИМ-метода) [7]. Согласно этому методу процесс планирования представляет собой развертывание конструктивного процесса пошагового формирования траектории движения из исходного состояния к целевому. Напомним, что при описании конструктивного процесса четко охарактеризован список исходных объектов, рассматриваемых в качестве неделимых, задан список правил образования новых объектов из ранее построенных, формирование новых объектов осуществляется из ранее построенных произвольным и, как правило, весьма простым способом в рамках сформулированных правил по шагам [8].

Лежащая в основе КИМ-метода идея заимствована из практики. Опытный руководитель, преследуя определенную цель и снабженный необходимой информацией принимает весьма эффективные оперативные решения и чем короче временной горизонт принимаемого решения, тем оно, как правило, лучше (решения, принимаемые «сегодня» относительно «завтрашнего дня», обычно эффективнее решений на события, отстоящие от настоящего времени на длительный период). Именно это обстоятельство часто обусловливает успех так называемого «ручного управления».

Ядром алгоритма формирования плана на основе КИМ-метода является осуществление одного элементарного шага на пути построения траектории движения объекта рассмотрения. Для этого на каждом шаге построения плана, исходя из сложившейся ситуации и в интересах поставленной цели формируется и оценивается (взвешивается) полное множество ресурсно и технологически допустимых частных решений

(действий). Из сформированного множества методом рандомизированного розыгрыша выбирается одно из частых решений, которое включается в качестве очередной составляющей будущего плана. Затем, согласно этому частному решению производится имитация его осуществления путем перевода объекта рассмотрения в новое фазовое состояние. Далее процедура планирования очередного шага повторяется, но уже из нового состояния системы и т.д.

Таким образом в процессе КИМ-метода объединены действия по планированию очередного шага и его реализации. По существу, реализуется симуляция целенаправленной пошаговой деятельности рассматриваемого объекта как бы в режиме ручного управления. В результате объединения процедур принятия частных решений и непосредственно следующих за этим процедур реализации последних получают вариант ресурсно и технологически допустимого (по построению) плана, как упорядоченной совокупности взаимосвязанных элементарных действий (уровень «элементарности» определяется разработчиком).

Сгенерировав достаточное количество вариантов, можно выбрать среди них наиболее подходящий и рассматривать его в качестве плана.

Опишем метод конструктивно-имитационного планирования в более формализованной виде. Пусть имеется возможность формального описания состояния рассматриваемого объекта *S,* как совокупности описаний каждого его компонента. Имеется множество представленных в формальном виде действий, рассматриваемых в качестве элементарных. Опыт применения КИМ-метода показал, что все эти описания удобно представлять в форме фреймов.

Пусть после осуществления $(k-1)$-го шага (т.е. к началу шага k) объект рассмотрения находится в конкретном состоянии S_i^k и требуется осуществить следующий k-ый шаг. Для каждого конкретного состояния S_i^k может быть сформировано множество потенциально возможных действий $\left\{d_r^k\right\}$.

Описывающие планы деятельности тексты в значительной мере носят регулятивный характер. В них описание каждого такого рода (элементарного) действия, как правило, представлено предложением, которое включает в себя описание глагола или глагольного выражения с описанием ряда требующихся при реализации этого действия обстоятельств, т.е. сирконстант $\left\{C_{rh}\right\}$, а также описание совокупности акторов этого действия $\left\{X_{ru}^k\right\}$ [9].

Реализация действия $d_{r^*}^k$ (r^* – фиксируется выбранное действие) заключается в переводе системы в новое состояние $S_j^k = \left\{Y_{r^*v}^k\right\}$, где $Y_{r^*v}^k$ описание состояния v–го компонента объекта рассмотрения по завершении шага k. Таким образом элементарное действие может рассматриваться в качестве оператора, переводящего состояние объекта рассмотрения вначале шага k в состояние в конце этого шага

$$d_{ij}^k : \; S_i^k \to S_j^k$$

Структура описывающего элементарное действие описывается стандартной структурой, представленной на Рис.1.

Выбранное действие d_{ij}^k, исходное состояние участвующих в этом действии акторов и результаты действия фиксируются в соответствующих фреймах. Описанная

процедура формирования частных решений в форме описанных «троек» повторяется до достижения поставленной цели, или фиксации невозможности ее достижения при данных условиях.

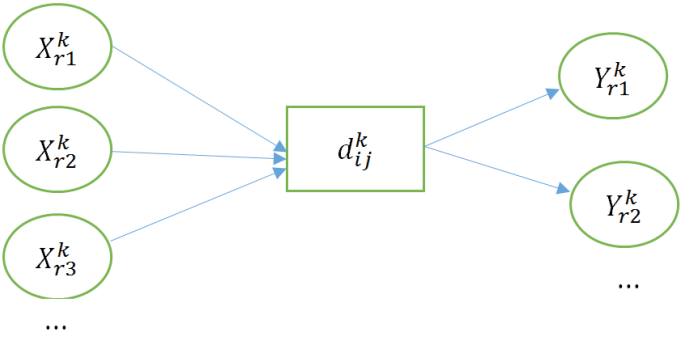

Рис.1. Структура модели элементарного этапа плана

После этого осуществляется этап сборки плана Π, как упорядоченного множество конкретных действий, каждый из которых описывает перевод из известного исходного состояния S_i в известное результирующее состояние S_j

$$\Pi = \{d_{ij}^k : S_i(k) \to S_j\ (k+1)\}$$

Важно отметить, что в разработанном плане подмножество результатов предыдущих действий составляет описание исходной ситуации для последующего действия. Наличие описанных троек позволяет в автоматическом режиме собрать единый граф, представляющий собой модель разработанного КИМ-плана. При этом описаны не только все реализующие план действия но и соответствующие промежуточные состояния всех задействованных в плане объектов

3. КИМ–РЕАЛИЗАЦИЯ

Применение метода КИМ-планирования обеспечивается представление разработанного плана в виде совокупности упорядоченных шагов, каждый из которых конструктивно формализован. Именно это обстоятельство позволяет разработать метод оценки вероятности реализации полученного плана, как конструктивную процедуру формирования его возможной реализации и построение соответствующей модели (КИМ-реализации).

Схема построения КИМ–реализации в значительной мере совпадает со схемой построения КИМ-плана. Однако в этом случае нет необходимости на каждом шаге формировать множество вариантов альтернативных частных решений и выбирать одно из них, поскольку для каждого шага оно уже присутствует в рассматриваемом плане (решается задача воспроизведение конкретного исследуемого план). Процедура же его пошаговой имитации остается. Поскольку она реализуется с учетом условий неопределенности, то получаемый результат каждого частного решения может

не совпадать с запланированным (он представляет собой реализацию соответствующего случайного процесса).

При имитации исполнения частного решения могут возникать самые различные ситуации, разобьём их на три группы:

1. В результат имитации рассматриваемого шага полученный результат практически соответствует плановому (его параметры находятся в допустимых пределах учитываемых показателей). В этом случае считают получаемый результат частного решения совпадающим с плановым и переходят к следующему согласно плану шагу.
2. Если в результате имитации параметры результатов частного решения выходят за допустимые пределы, но при этом имеется ресурсная и/или технологическая возможность компенсировать получившиеся рассогласования. Для таких случаев устанавливаются правила, согласно которым обеспечивается корректура данного частного решения с целью обеспечения выхода на плановое для начала следующего шага состояние за счет привлечения дополнительных ресурсов. Этот подход часто реализуется на практике, когда для обеспечения выхода на плановую траекторию разрешено вводить дополнительный ресурс, например, допускаются сверхурочные работы, использовать запасы нужных материалов и комплектующих.
3. Если в результате имитации запланированного частого решения получаются результаты, столь серьезно расходящиеся с плановыми, что выход на плановую траекторию в данной ситуации невозможен или не целесообразен, план считается нереализованным.

В результате прогона модели КИМ–реализации для данного КИМ-плана последний относится к реализованному (случай 1 и 2) или нереализованному (случай 3). Проведя серию прогонов модели КИМ-реализации и набрав достаточную статистику, может быть получена оценка вероятности его реализации.

Описанные процедуры выработки плана с определением значений результативных показателей и вероятности реализации плана могут быть положены в основу для разработки соответствующих алгоритмов и программных комплексов, обеспечивающих разрешение одной из проблем планирования сложной деятельности. Одна из схем их построения может быть следующей.

Пусть в результате имитации частного решения на k-м шаге получено состояние RS_j^k, которое, вообще говоря не совпадает с запланированным ($S_j^k \neq RS^k$). Формируется мера несовпадения состояний $\delta M^k = f\left(\Delta\left(X_{rv}^k, RS^k\right)\right)$. Обычно эта мера представляет собой функцию векторного аргумента, компоненты которого соответствуют фазовым координатам объекта рассмотрения.

Затем делается попытка для состояния RS_j^k реализовать последующее запланированное действие d_{ij}^{k+1}. В зависимости от объекта рассмотрения и проводимой его руководством политики задаются, например, двумя (в случае несимметричности показателей четыре) уровнями несовпадения плановых и фактических состояний: $\delta_1 M\left(d^k\right)$, $\delta_2 M\left(d^k\right)$, где:

$\delta_1 \bar{M}\left(d^k\right)$ – уровень нечувствительности: когда $\delta M^k \leq \delta_1 M\left(d^k\right)$, процесс проверки реализуемости продолжается;

$\delta_2 \bar{M}\left(d^k\right)$ – уровень неосуществимости плана: когда $\delta M^k > \delta_2 M\left(d^k\right)$. В этом случае процесс проверки реализуемости прерывается, о чем делается отметка в соответствующем разделе статистического материала.

В том случае, если $(\delta_1 M\left(d^k\right) \leq \delta M\left(d^k\right) \leq \delta_2 \bar{M}\left(d^k\right)$ – уровень нечувствительности превышен, но уровень неосуществимости плана не достигнут, необходима попытка введения допустимой коррекции условий. Такая коррекция осуществляется введением обусловленных ситуацией дополнительных ресурсов (обозначим их через ΔG^k).

Могут быть рассмотрены два варианта: запас дополнительных ресурсов лимитирован или считается всегда достаточным. Если запас дополнительных ресурсов зафиксирован $G0^k$, то после симуляции каждого частного решения, потребовавшего дополнительных ресурсов, вектор оставшегося дополнительного ресурса G^k обновляется ($G^{k-1} \to G^k$). В этой связи при необходимости использования дополнительных ресурсов осуществляется проверка возникшей ситуации:

Если $\Delta G^k > G^k$ (дополнительный ресурс в нужном объеме отсутствует), то данное частное решение не выполнимо и проверка реализуемости плана прекращается.

Если же $\Delta G^k \leq G^k$ (дополнительный ресурс имеется в нужном объеме), то фиксируется использование дополнительного ресурса и проверка реализуемости плана продолжается.

В том случае, если запас дополнительных ресурсов всегда считается достаточным, то всегда реализуется схема п. 2.

Таким образом после имитации рассматриваемого шага плана запускается процедура анализа сложившейся ситуации и принятие решения на способ продолжения КИМ-реализации. Возможны следующие альтернативы (см. рис. 2):

- процесс имитации продолжается, исходя из полученного за счет осуществления предыдущих шагов состояния;
- проводится процедура определения возможности возвращения сложившейся ситуации на плановую и по ее результатам принимаются меры по возвращению на плановую траектории;
- поскольку процедура определения возможности возвращения сложившейся ситуации приводит к отрицательному результату процесс проверки плана прекращается и фиксируется случай нереализации рассматриваемого плана.

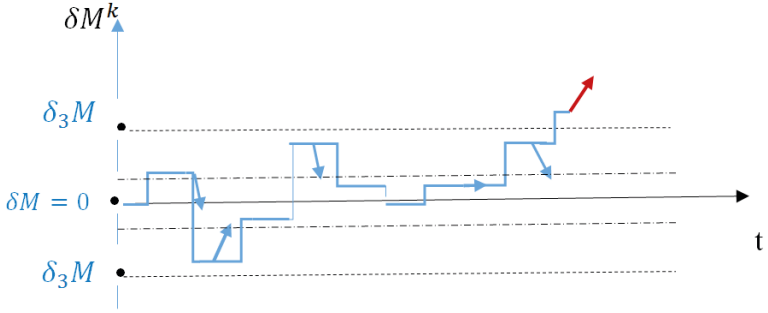

Рис. 2. Варианты частных решений при имитации процесса реализации плана

4. ЗАКЛЮЧЕНИЕ

Разработка на основе конструктивно-имитационного метода пары моделей КИМ-планирования и КИМ-реализации в силу их конструктивности возможно формирование программного комплекса, с помощью которого помимо собственно формирования плана обеспечивается возможность вести его анализ.

В частности, после одного прогона модели КИМ–реализации формируется детальная картина возможного осуществления разработанного плана, позволяющая выявлять причины возникновения различных ситуаций.

После осуществления серии прогонов модели появляется возможность получения статистических характеристик рассматриваемого плана, в частности:

- оценить вероятность реализации плана;
- выявить так называемые «узкие места» плана;
- определить ассортимент и объем ресурсов, которые целесообразно выделять для реализации данного плана;
- сформулировать предложения по корректуре структуры формирования плана и некоторых аспектов тактики и стратегии деятельности объекта рассмотрения.

Литература

1. *Бенвенисте, Гай.* Овладение политикой планирования: Создание реально выполнимых пл. и политики, которая ведет к переменам / Гай Бенвенисте; Пер. с англ. К.П. Михневич; Под общ. ред. М.Р. Калантаровой. М.: Прогресс: Фирма «Универс», 1994. 303 с.
2. Принятие решений при управлении организационными системами: Монография. / С.М. Вертешев [и др.] Псков: Псковский государственный университет, 2019. 218 с.
3. *Аладко А.В., Воронов М.В.* Модель оперативного планирования деятельности предприятия // Моделирование и анализ данных. 2016. – № 1. С. 37–47.
4. Анализ выполнимости планов мероприятий при оперативном управлении машиностроительным предприятием / Е.И. Шлычков [и др.] // Вестник Саратовского государственного технического университета. 2007. Вып. 1 (21). С. 88–95.
5. *Склемин А.А., Кушников В.А.* Анализ выполнимости планов мероприятий при управлении промышленным предприятием. // Известия высших учебных заведений. Поволжский регион. Технические науки. 2012. № 4 (24). С. 18–28.
6. *Шиянов Б.А. Силютина О.В., Неженец В.С.* Вероятностно-статистические методы количественной оценки рисков в системе регулирования неравновесными состояниями экономических систем // Вестник Воронежского государственного технического университета. 2010. № 8. С. 164–170.
7. *Воронов М.В.* Конструктивно-имитационное моделирование слабоструктурированных систем // Известия МАН ВШ. 2007. № 4(42). С.156–165.
8. Математическая энциклопедия. Т. 2. – М.: Советская энциклопедия, 1979.
9. *Воронов М.В.* Пименов В.И. Формализация регулятивных текстов // Информатика и автоматизация, 2021. Вып. 20, том 3. С. 562–590.

On Solving the Problem of Estimating the Probability of Fulfilling the Developed Plan

Mikhail V. Voronov*
Moscow state University of Psychology & Education (MSUPE), Moscow, Russia
ORCID: https://orcid.org/0000-0001-7839-6250
e-mail: mivoronov@yandex.ru

The issues of assessing the probability of the implementation of the developed enterprise activity plan are considered. The use of a constructive simulation method is proposed, which provides the possibility of developing software complexes for automatic development of plans with a full set of estimates of their characteristic indicators, including the probability of the plan implementation, on a single methodological basis.

Keywords: plan, constructive process, model, simulation, efficiency, probability.

For citation:
Voronov M.V. On Solving the Problem of Estimating the Probability of Fulfilling the Developed Plan. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*, 2022. Vol. 12, no. 1, pp. 16–26. DOI: https://doi.org/10.17759/mda.2022120102 (In Russ., abstr. in Engl.).

References

1. Benveniste, Gai. Ovladenie politikoi planirovaniya: Sozdanie real'no vypolnimykh pl. i politiki, kotoraya vedet k peremenam. Gai Benveniste; Per. s angl. K.P. Mikhnevich; Pod obshch. red. M.R. Kalantarovoi. M.: Progress: Firma "Univers", 1994. 303 p.
2. Prinyatie reshenii pri upravlenii organizatsionnymi sistemami: Monografiya. S.M. Verteshev [i dr.] Pskov: Pskovskii gosudarstvennyi universitet, 2019. 218 p.
3. Aladko A.V., Voronov M.V. Model' operativnogo planirovaniya deyatel'nosti predpriyatiya. Modelirovanie i analiz dannykh. 2016. – № 1. P. 37–47.
4. Analiz vypolnimosti planov meropriyatii pri operativnom upravlenii mashinostroitel'nym predpriyatiem / E.I. Shlychkov [i dr.] .Vestnik Saratovskogo gosudarstvennogo tekhnicheskogo universiteta. 2007. Vyp. 1(21).). P. 88–95.
5. Sklemin, A. A., Kushnikov V.A. Analiz vypolnimosti planov meropriyatii pri upravlenii promyshlennym predpriyatiem.. Izvestiya vysshikh uchebnykh zavedenii. Povolzhskii region. Tekhnicheskie nauki. 2012. № 4 (24). P. 18–28.
6. Shiyanov B.A. Silyutina O.V., Nezhenets V.S. Veroyatnostno-statisticheskie metody kolichestvennoi otsenki riskov v sisteme regulirovaniya neravnovesnymi sostoyaniyami ekonomicheskikh sistem. Vestnik Voronezhskogo gosudarstvennogo tekhnicheskogo universiteta. 2010. № 8. P. 164–170.
7. Voronov M.V. Konstruktivno-imitatsionnoe modelirovanie slabostrukturirovannykh system. Izvestiya MAN VSh. 2007. № 4(42). P. 156–165.

***Mikhail V. Voronov,** Doctor of Technical Sciences, Head of the Department of Applied Mathematics, Faculty of Information Technology, Moscow state University of Psychology & Education (MSUPE), Moscow, Russia, ORCID: https://orcid.org/0000-0001-7839-6250, e-mail: mivoronov@yandex.ru

8. Matematicheskaya entsiklopediya T. 2. – M.: Sovetskaya entsiklopediya, 1979.
9. Voronov, M.V. Pimenov V.I. Formalizatsiya regulyativnykh tekstov. Informatika i avtomatizatsiya, 2021. Vyp. 20, tom 3. P. 562–590.

Получена 10.01.2022
Принята в печать 31.01.2022

Received 10.01.2022
Accepted 31.01.2022

Моделирование и анализ данных
2022. Том 12. № 1. С. 27–48
DOI: https://doi.org/10.17759/mda.2022120103
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2022. Vol. 12, no. 1, pp. 27–48
DOI: https://doi.org/10.17759/mda.2022120103
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

◇◇◇◇◇◇◇ МЕТОДЫ ОПТИМИЗАЦИИ ◇◇◇◇◇◇◇

УДК 004.85

Выявление и классификация токсичных высказываний методами машинного обучения

Платонов Е.Н.*
Московский авиационный институт
(национальный исследовательский университет)
(ФГБОУ ВО МАИ), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0001-8502-1350
e-mail: en.platonov@gmail.com

Руденко В.Ю.**
Московский авиационный институт
(национальный исследовательский университет)
(ФГБОУ ВО МАИ), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0003-0010-331X
e-mail: super.ruden2011@mail.ru

Количество оставленных комментариев на платформах социальных сетей может составлять несколько миллионов в день, поэтому их владельцы заинтересованы в автоматической фильтрации контента. В данной работе рассматривается задача выявления оскорбительных высказываний в текстах. При решении задачи были рассмотрены различные метолы векторного преобразования текстов: TF-IDF, Word2Vec, GloVe и т.д. Так же были рассмотрены и представлены результаты применения классических методов классификации текста и нейросетевых методов (LSTM, CNN).

Ключевые слова: обработка текстов на естественном языке (NLP), задача классификации, градиентный бустинг, XGBoost, CatBoost, реккурентные нейронные сети, LSTM, сверточные нейронные сети.

***Платонов Евгений Николаевич,** кандидат физико-математических наук, доцент, Московский авиационный институт (национальный исследовательский университет) (ФГБОУ ВО МАИ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0001-8502-1350, e-mail: en.platonov@gmail.com

****Руденко Вероника Юрьевна,** студент магистратуры института «Информационные технологии и прикладная математика» Московский авиационный институт (национальный исследовательский университет), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0003-0010-331X, e-mail: super.ruden2011@mail.ru

Для цитаты:
Платонов Е.Н., Руденко В.Ю. Выявление и классификация токсичных высказываний методами машинного обучения// Моделирование и анализ данных. 2022. Том 12. № 1. С. 27–48. DOI: https://doi.org/10.17759/mda.2022120103

1. ВВЕДЕНИЕ

Обнаружение оскорбительного (токсичного) контента контента – одна из задач обработки естественного языка.

Обработка естественного языка (NLP) – это общее направление искусственного интеллекта и математической лингвистики. Она изучает проблемы компьютерного анализа и синтеза естественных языков и представляет собой огромный спектр задач разного уровня:

- распознавание текста, синтез речи;
- морфологический анализ, канонизация;
- синтаксический разбор, токенизация предложений;
- извлечение отношений, определение языка, анализ эмоциональной окраски и т.д. [1].

Проблема обнаружения токсичного контента является актуальной, так как платформы социальных сетей обеспечивают среду, в которой люди могут свободно участвовать в дискуссиях, узнавать о тенденциях, новостях и т.д.

Для частичного решения проблемы Google и Jigsaw выпустили экспериментальное расширение Tune, позволяющее управлять показом комментариев, которые пользователи видят в сети. Tune работает с Perspective API [2], который учится помечать негативные комментарии тысяч людей, помечая миллионы постов как спам, домогательства или непристойный контент. Как только комментарий определен как токсичный, Tune может настроить видимость таких комментариев.

В работах [3,4] рассматривается аналогичная задача классификации оскорбительных комментариев. Авторы работы [3] сравнивают различные подходы глубокого обучения к решению данной задачи на двух наборах данных: комментарии на страницах обсуждений Википедии и набор данных социальной сети Twitter. Были использованы такие архитектуры нейронных сетей как: сверточная нейронная сеть (CNN), рекуррентная нейронная сети на основе Gated Recurrent Unit (GRU), двунаправленная рекуррентная сеть (Bi-GRU). Авторы проводят детальный анализ ошибок первого (False Positive) и второго рода (False Negative). Причинами появления ошибок False Negative являются: токсичность предложения без использования ругательств, риторические вопросы, сарказм и ирония. Причинами появления ошибок False Positive являются: использование нецензурных слов в ложных срабатываниях, цитаты или ссылки.

Авторы работы [4] представляют различные подходы глубокого обучения, такие как CNN, LSTM, GRU. В качестве наборов данных были использованы общедоступные наборы данных, такие как Yahoo News Annotated Comments Corpus, комментарии

на страницах обсуждений Википедии, One Million Posts Corpus. Авторы использовали не двоичную, а мультиклассовую классификацию. Больше классов требует больше данных для обучения. Обычно для англоязычных текстов доступны большие объемы обучающих данных. Тем не менее, для менее распространенных языков обучающие данные редки, а иногда маркированные данные полностью отсутствуют. Одним из способов решения этой проблемы является машинный перевод англоязычного набора данных на другой язык. Если машинный перевод хорошего качества, аннотации англоязычных комментариев также применяются к переведенным комментариям. Поэтому для подобного расширения учебных данных были использованы исследования в области трансферного обучения [5].

В статье [6] было проведено исследование аналогичной задачи для корпусов текстов на русском и украинском языках. Основная идея представленного подхода состоит в том, чтобы использовать начальный словарь оскорбительных терминов в сочетании с неконтролируемым присвоением меток (оскорбительных или не оскорбительных) комментариям в социальных сетях, а затем итеративно его расширять оскорбительными словами.

В нашей работе задача обнаружение оскорбительного контента рассматривается как задача бинарной классификации.

2. ВЕКТОРНОЕ ПРЕДСТАВЛЕНИЕ ТЕКСТА

Векторное представление текста (word embedding) – это собирательное название для набора методов моделирования и обработки естественного языка (NLP), где слова или фразы из словаря отображаются в пространство действительных чисел.

Векторная модель – представление текстов вещественными векторами из одного общего для всей коллекции текстов векторного пространства. с фиксированной размерностью Размерность пространства равна количеству различных слов во всей текстовой коллекции.

Вектор образуется упорядочением весов всех слов, включая те, которых нет в конкретном тексте. Размерность этого вектора, как и размерность пространства, равна количеству различных слов во всей коллекции, и является одинаковой для всех текстов коллекции.

Формально:

$$d_j = (w_{j1}, w_{j2}, \dots, w_{jn})$$

где d_j – векторное представление -го текста, w_{ji} – вес i-го слова в j-м тексте, n – общее количество различных слов во всех текстах коллекции.

Для полного определения векторной модели необходимо указать, каким именно образом будет отыскиваться вес слова в документе [7].

Векторное представление текста является основным способом решением задач информационного поиска: поиск документа по запросу, классификация документов, кластеризация документов и т.д.

Рассмотрим несколько методов векторного представления слов.

TF-IDF

В данном методе текст сводится к вектору, где каждая его компонента представляет собой слово, а значением данной компоненты является число раз, которое это слово используется в тексте.

Преобразование TF-IDF – используется для корректировки значений вектора в соответствии с числом документов, использующих слово. Слова, встречающиеся во многих документах, могут быть менее значимыми, чем слова, встречающиеся реже. TF-IDF уменьшает значение данного слова пропорционально количеству документов, в которых оно появляется.

TF или частота слова – это отношение количества вхождения конкретного термина к суммарному набору слов в исследуемом тексте или же документе.

IDF или обратная частота документа – это инверсия частотности, с которой определенное слово фигурирует в коллекции текстов [8].

Word2Vec

Word2Vec – технология от Google [9], использующаяся для статистического анализа больших массивов текстовой информации. Она собирает статистику по совместному появлению слов в фразах, после этого с помощью нейронных сетей решает задачу уменьшения размерности и в итоге выдает компактные векторные представления слов, достаточно полно отражающие отношения этих слов в обрабатываемых текстах.

Данная модель предсказывает вероятность слова по его окружению – контексту. То есть формируются такие вектора слов, чтобы вероятность, присваиваемая моделью слову, была близка к вероятности встретить слово в этом окружении в реальном тексте [10].

FastText

FastText – это библиотека для изучения встраивания слов и классификации текста, созданная исследовательской лабораторией AI в Facebook. Модель позволяет создать алгоритм обучения без контроля или обучения для получения векторных представлений для слов.

Для эффективной обработки массивов данных с большим количеством различных категорий FastText использует иерархический классификатор, который организует категории в древовидную структуру вместо плоской.

FastText является расширением Word2Vec. Для модели векторных представлений слов используется *skip-gram* с негативным сэмплированием.

Негативное сэмплирование – это способ создать для обучения векторной модели отрицательные примеры, то есть показать ей пары слов, которые не являются соседями по контексту (например, «пушистый котик» → «пушистый утюг»). Всего подбирается от 3 до 20 случайных слов. Такой случайный подбор нескольких примеров не требует много компьютерного времени и позволяет ускорить работу FastText [11].

GloVe

GloVe (Global Vectors) – модель, предложенная лабораторией компьютерной лингвистики Стенфордского университета. Данный алгоритм сочетает в себе черты

SVD разложения и Word2Vec. Метод GloVe основан на идее выведения семантических отношений между словами из матрицы совпадений.

По входному словарю V, строится частотная матрица совпадений V x V, где элемент матрицы X_{ij} обозначает, сколько раз слово i встречалось со словом j.

	the	cat	sat	on	mat
the	0	1	0	1	1
cat	1	0	1	0	0
sat	0	1	0	1	0
on	1	0	1	0	0
mat	1	0	0	0	0

Рис. 1. Пример матрицы совпадений

Более подробно работа GloVe описана в статье [12].

Paragram

Paraphrastic sentence embeddings (Paragram) — универсальный эмбеддинг, основанный на комбинировании вложения слов для получения вложений предложений, удовлетворяющих тому свойству, что вектора предложений, являющиеся парафразами друг друга, расположены рядом друг с другом в векторном пространстве.

Парафраз – выражение, являющееся описательной передачей смысла другого выражения или слова.

Комбинирование проводится контролируемым образом на основе наблюдения из базы данных известных парафразов.

Цель алгоритма состоит в том, чтобы встроить последовательности в низкоразмерное пространство таким образом, чтобы косинусное сходство в пространстве соответствовало силе связи парафразирования между последовательностями [13].

3. МЕТОДЫ

В работе используются методы такие как логистическая регрессия, градиентный бустинг (XGBoost, CatBoost), нейронная сеть долговременной краткосрочной памяти (LSTM) и сверточная нейронная сеть (CNN).

Градиентный бустинг

Градиентный бустинг — техника машинного обучения для задач классификации и регрессии, которая строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений.

Обучение ансамбля проводится последовательно. На каждой итерации вычисляются отклонения предсказаний уже обученного ансамбля на обучающей выборке.

Следующая модель, которая будет добавлена в ансамбль будет стараться уменьшить эти отклонения. Таким образом, добавив предсказания нового дерева к предсказаниям обученного ансамбля происходит уменьшение среднего отклонения модели. Новые деревья добавляются в ансамбль до тех пор, пока ошибка уменьшается, либо пока не выполняется одно из правил «ранней остановки».

Метод градиентного бустинга обладает высокой гибкостью для решения задач классификации.

В данной работе будет использовано две реализации градиентного бустинга:

- XGBoost (eXtreme Gradient Boosting);

В отличие от стандартного градиентного бустинга в методе XGBoost построение деревьев основано на параллелизации, а в качестве критерия остановки разбиения дерева используется параметр максимальной глубины. Для нахождения оптимальных точек разделения используется метод взвешенных квантилей [14]. Так же алгоритм содержит возможность добавления L1 или L2 регуляризацию.

Еще одним значительным улучшением XGBoost является возможность упрощенной работы с разреженными матрицами.

Алгоритм работы XGBoost подробно описан в статье [14].

- CatBoost (Categorical Boosting).

Библиотека CatBoost (Categorical Boosting) – метод машинного обучения, основанный на градиентном бустинге, разработанный инженерами Яндекса. Главное преимущество которой заключается в том, что она одинаково хорошо работает как с числовыми признаками, так и с категориальными.

CatBoost основан на двоичных деревьях решений с градиентным бустингом. Во время обучения набор деревьев решений строится последовательно. Каждое последующее дерево строится с меньшими потерями по сравнению с предыдущими деревьями. Более подробно алгоритм и преимущество использования открытой библиотеки CatBoost описано в работе [15]. В статье также приведены результаты сравнения метода CatBoost с другими методами, использующими градиентный бустинг.

LSTM

Сеть долговременной краткосрочной памяти LSTM (Long Short-Term Memory) – частный случай рекуррентной нейронной сети (RNN).

RNN представляет собой сети с петлями в них, что позволяет хранить информацию о том, что было в предшествующем предложении. RNN используют обратное распространение ошибки в процессе обучения для обновления весов сети на каждом уровне.

Сеть долговременной кратковременной памяти была изобретена с целью решения проблем исчезающих и взрывающихся градиентов. Ключевым моментом в разработке LSTM было включение нелинейных, зависящих от данных элементов управления в ячейку RNN, которые могут быть обучены для обеспечения того, чтобы градиент целевой функции по отношению к сигналу состояния не исчезал [16].

Рассмотрим модуль LSTM, называемый блоком памяти, на основе работ [17, 18].

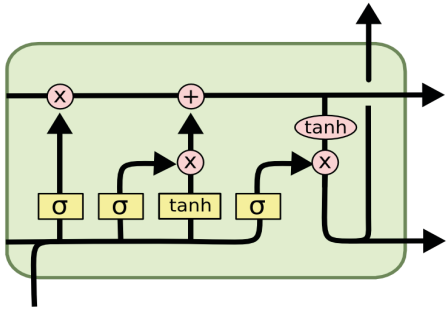

Рис. 2. Блок памяти

Входные затворы (gates), которые представляют собой простые сигмоидальные пороговые блоки с диапазоном функций активации [0, 1], управляют сигналами от сети к ячейке памяти, соответствующим образом масштабируя их; когда затвор закрыт, активация близка к нулю. Сигмоидальные затворы состоят из слоя сигмовидной нейронной сети и операции точечного умножения. Кроме того, они могут научиться защищать содержимое от помех со стороны неуместных сигналов.

Выходные затворы могут научиться управлять доступом к содержимому ячейки памяти, которое защищает другие ячейки памяти от помех. Выходной затвор решает какую информацию от предыдущих шагов необходимо сохранить. Результат будет являться отфильтрованным состоянием ячейки. Сначала текущий ввод проходит через сигмоидальный функцию, затем пропускаем состояние ячейки через функцию гиперболического тангенса (*tanh*). Вывод функции *tanh* перемножается на вывод сигмоидальной функции.

CNN

Сверточная нейронная сеть (Convolutional Neural Network, CNN) – архитектура нейронных сетей, как аналог зрительной коры головного мозга для распознавания изображений.

Структура сети – однонаправленная, принципиально многослойная. Для обучения используются стандартные методы, чаще всего метод обратного распространения ошибки.

Классическая архитектура CNN обычно имеет 3 типа слоев:

1. Сверточный слой.
2. Слой субдискретизации (pooling).
3. Полносвязный слой.

Сверточный слой получает в качестве входных данных трехмерную матрицу размера $[n_1, m_1, d_1]$.

Ядро (фильтр) сверточного слоя (сверточная матрица) представляет собой матрицу с размерами $[n_2, m_2, d_1]$, т.е. глубина d_1 ввода и одного ядра одинаковы. Для каждого сверточного слоя имеется несколько ядер, уложенных друг на друга, что образует матрицу с размерами $[n_2, m_2, d_2]$, где d_2 – количество ядер. Для каждого ядра есть соответствующее смещение, которое является скалярной величиной.

Выходом сверточного слоя является матрица размерностью $[n_3, m_3, d_2]$, глубина вывода равна количеству ядер.

Основная цель слоя субдискретизации состоит в том, чтобы уменьшить количество параметров входа. Входные данные для данного слоя являются тензорными.

Полносвязный слой – это просто нейронная сеть с обратной связью. Полносвязные слои образуют последние несколько слоев в сети. После прохождения через них последний слой использует функцию активации softmax, которая используется для получения вероятности того, что входные данные относятся к определенному классу [19].

Было показано, что сверточные нейронные сети, первоначально изобретенные для компьютерного зрения, обеспечивают высокую производительность при решении задач классификации текста [20].

В настоящее время предполагается, что CNN классифицирует текст, выполняя следующие шаги.

1. В качестве детекторов n-граммы используются одномерные сверточные матрицы, каждая из которых специализируется на тесно связанном семействе n-граммов. Ядра сверточного слоя не являются однородными, т. е. одно ядро может и часто обнаруживает несколько явно разных семейств n-грамм. Ядра также обнаруживают отрицательные элементы в n-граммах.
2. Слой субдискретизации с функцией максимума (Max Pooling) с течением времени извлекает соответствующие n-граммы для принятия решения. Max Pooling так же вызывает пороговое поведение, и значения ниже заданного порога игнорируются при прогнозировании.
3. Остальная часть сети классифицирует текст на основе этой информации.

4. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Рассмотрим задачу бинарной классификации текстов на примере конкурса от компании Quora с платформы Kaggle), в котором необходимо предсказать является ли вопрос, заданный на платформе *Quora* «искренним» – нейтральным или «неискренним» – токсичным [21].

Файл данных включает в себя 80810 неискренних (токсичных) записей и 1225312 нейтральных записей. Следовательно, при использовании методов решения задачи классификации следует учитывать тот факт, что работа будет вестись с несбалансированным набором данных.

Приведем часть данных в виде таблицы 1.

Таблица 1

Данные из набора Quora

	question text	**target**
0	How did Quebec nationalists see their province as a nation in the 1960s?	0
1	Do you have an adopted dog, how would you encourage people to adopt and not shop?	0
22	Has the United States become the largest dictatorship in the world?	1

	question text	**target**
…	…	…
1306093	How is it to have intimate relation with your cousin?	1
1306112	Are you ashamed of being an Indian?	1

Качество решения задачи классификации будем оценивать с помощью матрицы ошибок.

Таблица 2

Матрица ошибок

	$a = 0$	$a = 1$
$y = 0$	TP	FP
$y = 1$	FN	TN

Два класса делятся на положительный (обычно метка 1) и отрицательный (обычно метка 0). Объекты, которые алгоритм относит к положительному классу, называются положительными (Positive), те из них, которые на самом деле принадлежат к этому классу – истинно положительными (True Positive), остальные – ложно положительными (False Positive). Аналогичная терминология есть для отрицательного (Negative) класса. В табл. 2 использованы сокращения: TP = true positive, TN = true negative, FP = false positive, TN = true negative.

При построении матрицы ошибок также используются нормированные значения.

Таблица 3

Нормированная матрица ошибок

	$a = 0$	$a = 1$
$y = 0$	$\frac{TP}{TP + FN}$	$\frac{FN}{FN + TP}$
$y = 1$	$\frac{FP}{FP + TN}$	$\frac{TN}{TN + FP}$

Часто результат работы алгоритма на фиксированной тестовой выборке визуализируют с помощью ROC-кривой (receiver operating characteristic) [22], а качество оценивают как площадь под этой кривой – AUC (area under the curve). AUC ROC равен доле пар объектов вида (объект класса 1, объект класса 0), которые алгоритм верно упорядочил, т.е. первый объект идёт в упорядоченном списке раньше.

Применение классических методов

Рассмотрим решение поставленной задачи с помощью базовых алгоритмов классификации: логистической регрессии [23], XGBoost и CatBoost.

В этом разделе в качестве эмбединга будем использовать TF-IDF с размерностью выходного вектора равной 300.

В данных присутствует сильный дисбаланс классов (большой перевес в сторону нейтральных записей), это значит, что без настройки модели будут хорошо предсказывать нейтральный записи и плохо токсичные, при этом значении метрики качества будет высоким. Суть задачи же заключается как раз в поиске токсичных записей.

Для работы с несбалансированными данными есть несколько подходов. Одним из них является указание весов для каждого класса. Для каждой модели значение весов указывается по-разному.

Для инициализации метода логистической регрессии был использован параметр $class\_weight = 'balanced'$. Метол логистической регрессии при данном параметре использует значения целевой переменной y для автоматической регулировки весов, обратно пропорциональных частотам классов во входных данных.

Построим ROC-кривую и матрицу ошибок для метода логистической регрессии.

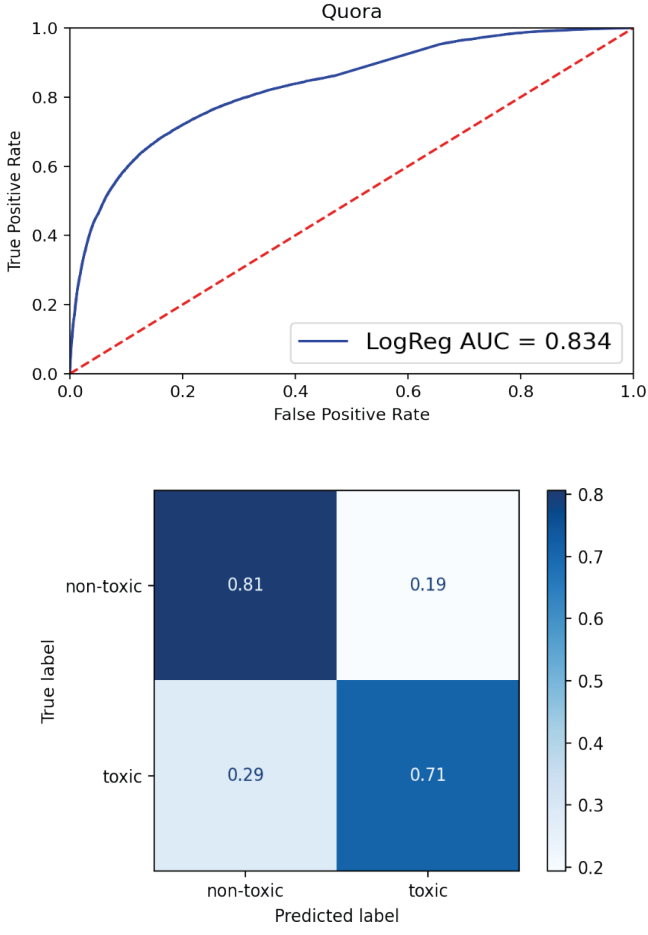

Рис. 4. ROC-кривая и матрица ошибок для логистической регрессии

Выведем топ-10 весов слов для каждого класса с помощью библиотеки для интерпретации результатов eli5 [24].

y=toxic top features

Weight[?]	Feature
+4.975	muslims
+4.437	indians
+4.271	americans
+4.071	trump
+3.683	muslim
+3.524	women
+3.353	girls
+3.086	white
+2.930	hate
+2.863	sex
… 150 more positive …	
… 131 more negative …	
-2.279	best
-2.384	rank
-2.407	energy
-2.429	app
-2.516	prepare
-2.683	tips
-2.762	marketing
-2.978	engineer
-3.084	marks
-3.101	engineering

Рис. 5. Топ-10 весов слов для каждого класса

Зеленым выделены веса слов, по которым метод логистической регрессии относит текст к классу токсичный или неискренний, красным соответственно веса слов для класса нейтральных записей. Например, если в вопросе присутствует слово «muslim», то он с большей вероятностью будет отнесен к классу токсичных вопросов.

При инициализации метода CatBoost были переданы веса классов: $class_{weights} = [0.06, 0.94]$. Веса классов для данной модели были вычислены как отношение:

$$count_0 / n; count_1 / n,$$

где $count_0$ – количество нейтральных записей в обучающей выборке, $count_1$ – количество оскорбительных записей в обучающей выборке, n – размер обучающей выборки.

Для инициализации метода XGBoost был рассчитано отношение записей на обучающем наборе нейтрального класса к классу токсичных вопросов: $estimate = 15.160$.

В метод был передан параметр $scale\_pos\_weight = (estimate + 10) = 25.160$. Добавка $+10$ была подобрана эмпирически. Параметр отвечает за балансировку положительных и отрицательных весов.

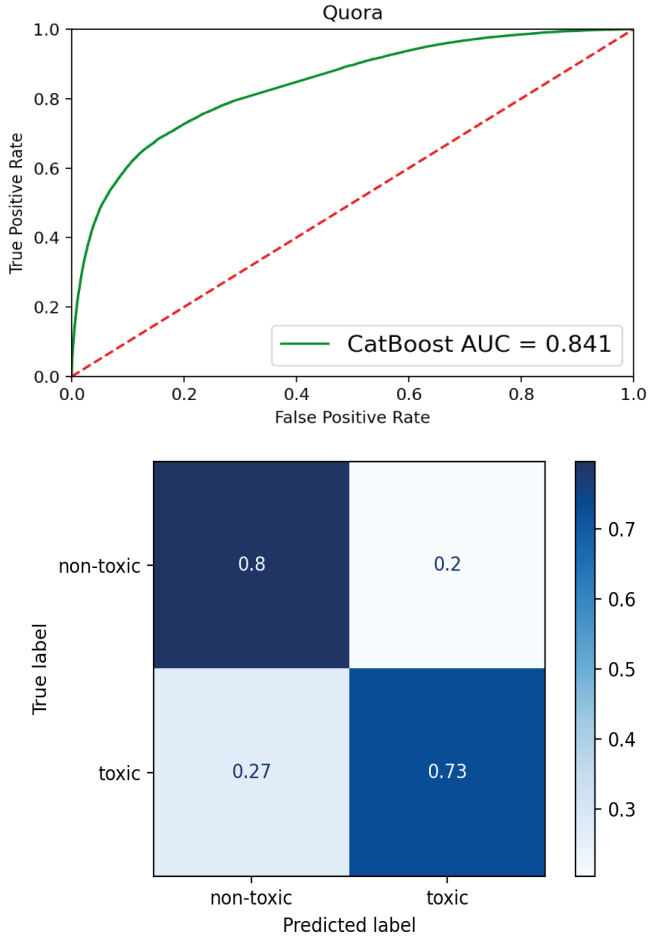

Рис. 6. ROC-кривая и матрица ошибок для CatBoost

В примере с логистической регрессией были показаны топ 10 весов для каждого класса, которые вносят наибольший вклад при классификации текстов. Для дальнейших примеров данную таблицу выводить не будем и будем рассматривать интерпретацию отдельного текста из коллекции с помощью алгоритма LIME для модели XGBoost.

LIME (Local Interpretable Model Agnostic Explanations) – алгоритм, который может объяснить предсказания классификатора или регрессора, локально аппроксимируя его интерпретируемой моделью [25].

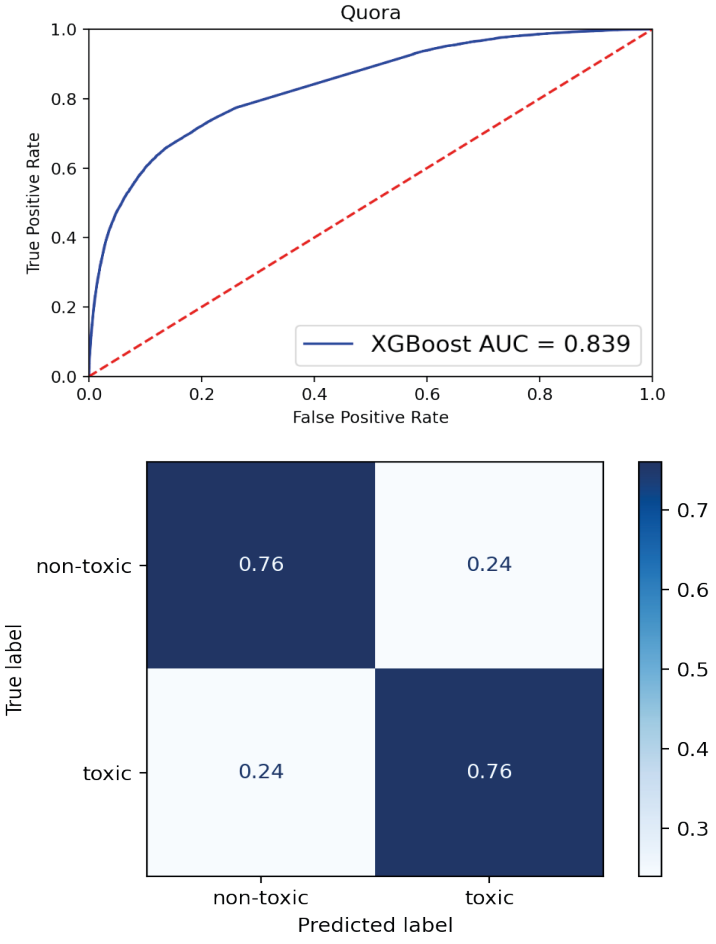

Рис. 7. ROC-кривая и матрица ошибок для XGBoost

Благодаря весу слова «working» текст был определен как нейтральный. Веса всех слов кроме «working» достаточно малы и поэтому на представленном рисунке отображаются со значениями 0.00. Алгоритм LIME позволяет отдельно вывести все веса слов и объяснить их значения для каждого класса. Так как значение данного класса равно 0 выведем веса слов для него:

$$Explanation\ for\ class\ 0$$
$$('working', 0.0732)$$
$$('the', 8.8608e^{-6})$$
$$('of', 8.8331e^{-6})$$
$$('stands', 8.0563e^{-6})$$

$('What', 6.6930e^{-6})$
$('potentiometer', 6.4858e^{-6})$
$('which', 5.9871e^{-6})$
$('basics', 4.7359e^{-6})$
$('upon', 4.3065e^{-6})$
$('are', 3.5078e^{-6})$

Так же получим результаты для Logistic Regression, XGBoost и CatBoost с применением эмбедингов Word2Vec и FastText.

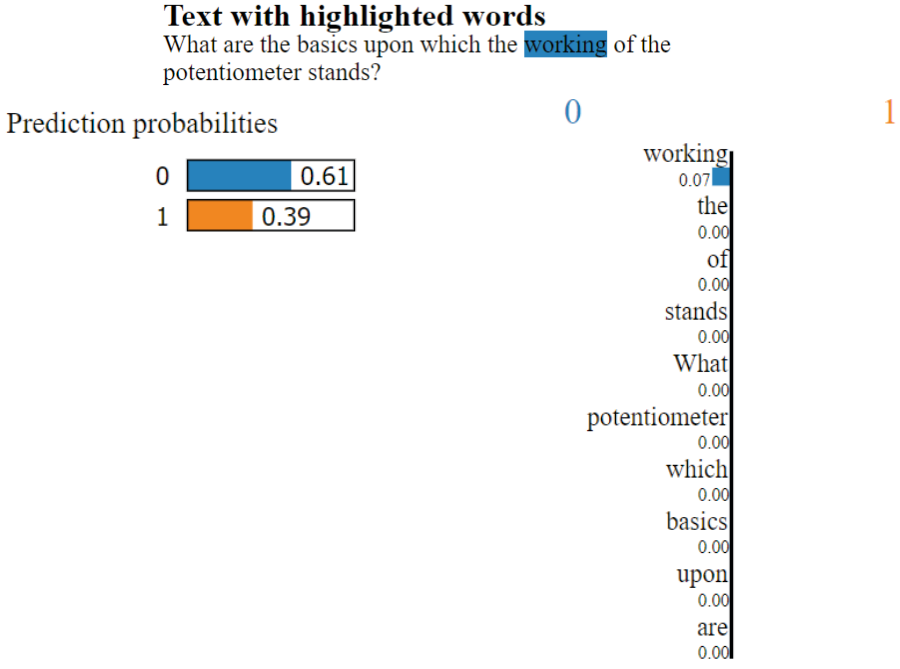

Рис. 8. Интерпретация результатов с помощью LIME

Таблица 4

Сводная таблица точности AUC

	Word2Vec		TF-IDF		FastTex	
	ROC-AUC	**Time**	**ROC-AUC**	**Time**	**ROC-AUC**	**Time**
LogReg	0.668	0.81	0.835	0.17	0.665	0.76
CatBoost	0.761	10.15	0.839	12.08	0.759	9.59
XGBoost	0.751	30.24	0.841	0.38	-	-

Из полученных результатов видно, что при использовании TF-IDF стандартные модели показывают более высокое значение метрики качества.

Применение методов глубокого обучения

Чтобы сократить время обучения моделей из этого раздела и подготовки текста для векторного представления был использован предобученный Embedding слой.

Предобученные эмбеддинги обычно представляют собой словари, где слову ставится в соответствие вектор. Такие словари получаются следующим образом: эмбеддинг обучается на больших корпусах текстов с использованием специальных моделей, а затем выгружается в виде словаря. Таким образом можно получить качественный эмбеддинг без необходимости его повторного обучения.

Зададим размер словаря передаваемый в Embedding слой равным 50000 уникальных слов. Для покрытия большинства примеров достаточно взять длину предложения равную 20 словам (рис. 9).

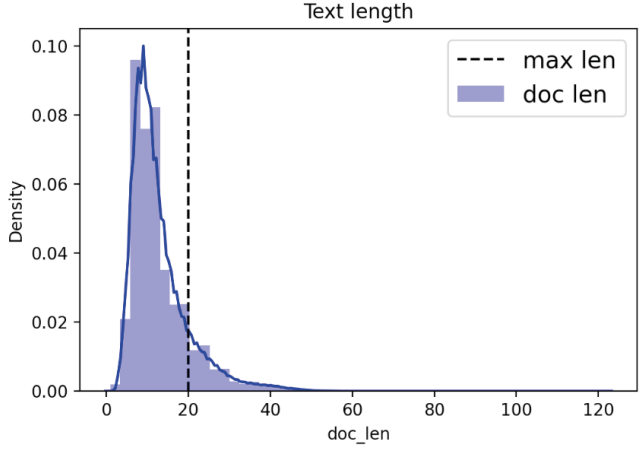

Рис. 9. Гистограмма длин предложений в обучающей выборке

В аргумент weights Embedding слоя передается эмбеддинг матрица, составленная из четырех предобученных эмбеддингов: GloVe [26], FastText [27], Paragram [28] и Word2Vec [29]. Использование ансамбля предобученных эмбеддингов позволит моделям более точно классифицировать тексты.

Для каждого из представленных эмбедингов выходной вектор имеет размерность 300, как и в случае с TF-IDF из прошлого раздела. Тогда выходная размерность Embedding слоя равна $300 \times 4 = 1200$.

Для моделей глубокого обучения вычислим веса классов как:

$$n\_samples \,/\left(n_{classes} * bincount\left(y\right)\right),$$

где $n\_samples$ – размер выборки, $n_{classes}$ – количество классов (в нашем случае 2), $bincount\left(y\right)$ – функция, подсчитывающая количество вхождений каждого класса в выборку.

В таком случае веса классов равны: $\left\{0:0.533, 1:8.062\right\}$.

Рассмотрим решения получаемы с использованием LSTM и CNN.

Сеть LSTM была обучена с batch_size = 1024 и количеством эпох равном 5. Построим ROC-кривую и матрицу ошибок.

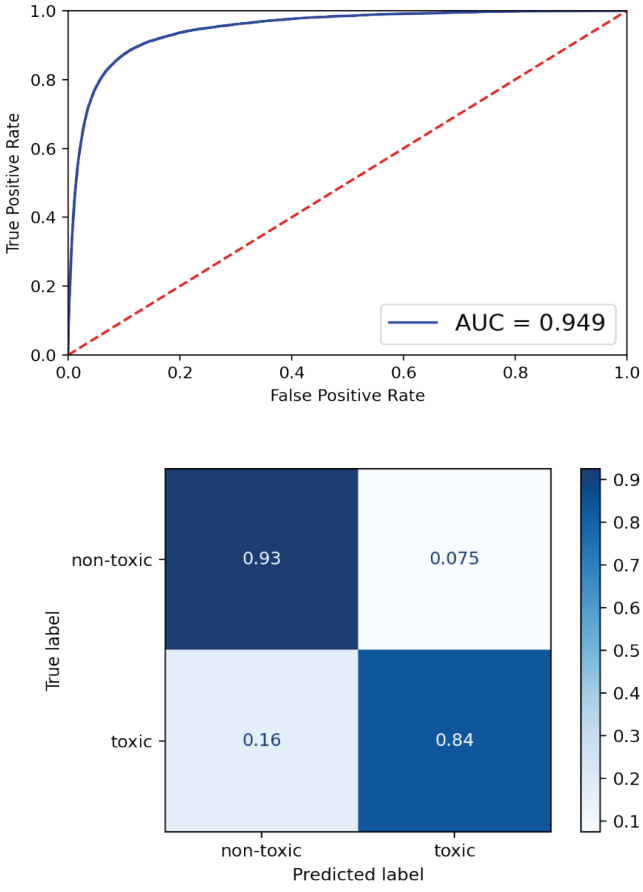

Рис. 10. ROC-кривая и матрица ошибок для LSTM

Сеть CNN была обучена с batch_size = 512 и двумя эпохами обучения. Построим ROC-кривую и матрицу ошибок.

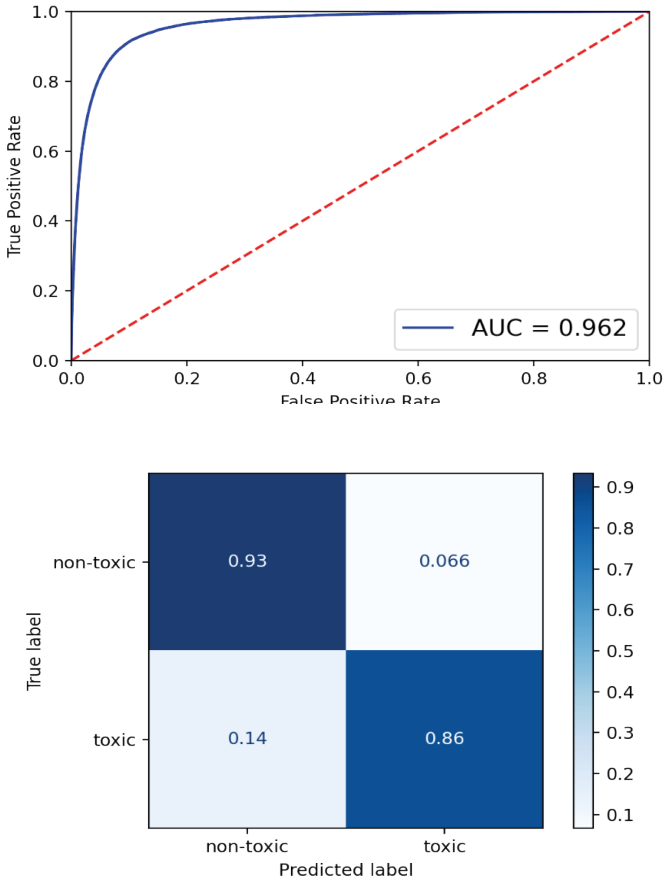

Рис. 11. ROC-кривая и матрица ошибок для CNN

Проинтерпретируем результаты архитектуры LSTM для того же предложения с помощью алгоритма LIME.

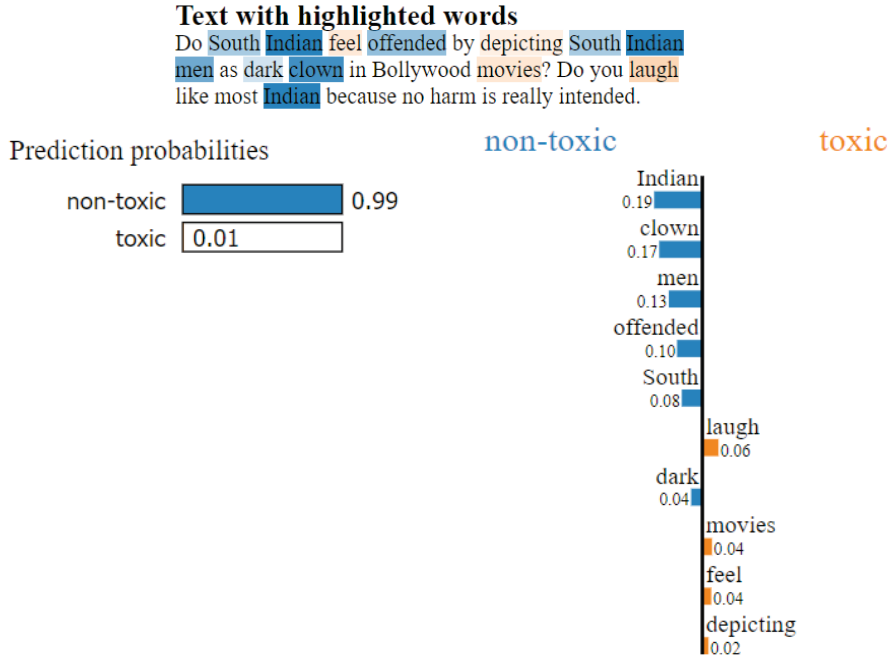

Рис.12. Интерпретация результатов с помощью LIME

На рис. 12 можно увидеть благодаря весам каких слов данный текст был отмечен как нейтральный.

4. ЗАКЛЮЧЕНИЕ

Сравним результаты, полученные с помощью моделей глубоко обучения, с классическими моделями решения задачи классификации.

Таблица 5

Сводная таблица точности AUC и элементов матрицы

Применяемые методы	ROC-AUC	balanced accuracy
LogReg	0.835	0.760
CatBoost	0.839	0.765
XGBoost	0.841	0.760
LSTM	0.949	0.875
CNN	0.962	0.895

Balanced accuracy – показатель, который можно использовать при оценке того, насколько хорош, бинарный классификатор:

$$Balanced\ accuracy = \frac{1}{2}\left(\frac{TP}{TP+FP}+\frac{TN}{TN+FP}\right).$$

Видно, что CNN превосходит как классические модели, так и LSTM по значению метрики качества и по значению balanced accuracy.

Результаты для обеих моделей LSTM и CNN были получены с использованием малого числа эпох, когда как для количества тренируемых параметров, значение которых достигает нескольких миллионов, этого явно недостаточно. Так же из-за факта обучения на малом количестве эпох вытекает необходимость исследования на предмет возникновения эффекта переобучения моделей.

При дальнейшей работе с бинарной классификацией текстов имеет смысл обратить внимание на такие модели как Bi-LSTM, GRU и Bi-GRU, на применение методов ансамблирования и трансферного обучения (модели BERT, ELMo и т.д.).

Литература

1. *Риз Р.* Обработка естественного языка на Java. ДМК-Пресс. 2016. 264 с.
2. Perspective API. URL: https://www.perspectiveapi.com
3. *van Aken B., Risch J., Krestel R., Löser A.* Challenges for toxic comment classification: An in-depth error analysis. 2018, arXiv:1809.07572.
4. *Risch J., Krestel R.* Toxic Comment Detection in Online Discussions. Deep Learning-Based Approaches for Sentiment Analysis. Springer, Singapore, 2020. P. 85–109.
5. *Weiss K., Khoshgoftaar T.M., Wang D.* A survey of transfer learning // Big Data, 3: 9. 2016. https://doi.org/10.1186/s40537–016–0043–6
6. *Andrusyak B., Rimel M., Kern R.* Detection of Abusive Speech for Mixed Sociolects of Russian and Ukrainian Languages // RASLAN. – 2018. – P. 77–84.
7. *Li Y., Yang T.* Word Embedding for Understanding Natural Language: A Survey. In: Srinivasan S. (eds) Guide to Big Data Applications. Studies in Big Data, vol 26. Springer, Cham. https://doi.org/10.1007/978–3-319–53817–4_4
8. *Liu C.* et al. Research of text classification based on improved TF-IDF algorithm // IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE). 2018 P. 218–222.
9. word2vec // URL: https://code.google.com/archive/p/word2vec/
10. *Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean.* Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR, 2013.
11. *Bojanowski P*, et al. Enriching word vectors with subword information // Transactions of the Association for Computational Linguistics. 2017. V. 5. P. 135–146.
12. *Pennington J., Socher R., Manning C.D. G*love: Global vectors for word representation // Proceedings of the conference on empirical methods in natural language processing (EMNLP). 2014. P. 1532–1543.
13. *Wieting J.* et al. From paraphrase database to compositional paraphrase model and back // Transactions of the Association for Computational Linguistics. 2015. V. 3. P. 345–358.
14. *Chen T., Guestrin C.* Xgboost: A scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. P. 785–794.
15. *Dorogush A.V., Ershov V., Gulin A*. CatBoost: gradient boosting with categorical features support // arXiv preprint arXiv:1810.11363. 2018.
16. *Sepp Hochreiter and Jürgen Schmidhuber.* Long Short-Term memory // Neural computation, V. 9(8). P. 1735–1780, 1997.
17. *Staudemeyer R.C., Morris E.R.* Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks // arXiv preprint arXiv:1909.09586. 2019. URL:https://arxiv.org/pdf/1909.09586.pdf
18. Understanding LSTM Networks URL:https://colah.github.io/posts/2015–08-Understanding-LSTMs/
19. Convolutional Neural Network. An Introduction to Convolutional Neural Networks. URL: https://towardsdatascience.com/convolutional-neural-network-17fb77e76c05

20. *Bai S., Kolter J.Z., Koltun V.* An empirical evaluation of generic convolutional and recurrent networks for sequence modeling // CoRR, abs/1803.01271. 2018. http://arxiv.org/abs/1803.01271
21. Quora Insincere Questions Classification. URL: https://www.kaggle.com/c/quora-insincere-questions-classification/data
22. *T. Fawcett.* An introduction to ROC analysis // Pattern Recognition Letters, V. 27. 2006. P. 861–874.
23. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Springer-Verlag, New York. 2017.
24. Eli5 Documentation. URL: https://eli5.readthedocs.io/en/latest/
25. *Tulio Ribeiro M., Singh S., Guestrin C.* Why Should I Trust You? Explaining the Predictions of Any Classifier // KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 1135–1144.
26. glove.840B.300d – pre-trained word vectors GloVe. URL: https://nlp.stanford.edu/projects/glove/
27. wiki-news-300d-1M – pre-trained word vectors trained using fastText. URL: https://fasttext.cc/docs/en/english-vectors.html
28. paragram_300_sl999 – New Paragram-SL999 300 dimensional embeddings tuned on SimLex999 dataset. URL: https://www.kaggle.com/ranik40/paragram-300-sl999
29. GoogleNews-vectors-negative300 – pre-trained word vectors trained using Word2Vec. URL: https://code.google.com/archive/p/word2vec/

Identification and Classification of Toxic Statements by Machine Learning Methods

Evgeniy N. Platonov*
Moscow Aviation Institute (National Research University) Moscow, Russia
ORCID: https://orcid.org/0000-0001-8502-1350
e-mail: en.platonov@gmail.com

Veronika Y. Rudenko**
Moscow Aviation Institute (National Research University) Moscow, Russia
ORCID: https://orcid.org/0000-0003-0010-331X
e-mail: super.ruden2011@mail.ru

The number of comments left on social media platforms can reach several million per day, so their owners are interested in automatic content filtering. In this paper, the task of identifying offensive statements in texts is considered. When solving the problem, various methods of vector text conversion were considered: TF-IDF, Word2Vec, Glove, etc. The results of the application of classical text classification methods and neural network methods (LSTM, CNN) were also considered and presented.

Keywords: Natural Language Processing (NLP), Classification, Gradient boosting, XGBoost, CatBoost, Recurrent Neural Network, LSTM, Convolutional Neural Network.

For citation:
Platonov E.N., Rudenko V.Y. Identification and Classification of Toxic Statements by Machine Learning Methods. *Modelirovanie i analiz dannykh = Modelling and Data Analysis,* 2022. Vol. 12, no. 1, pp. 27–48. DOI: https://doi.org/10.17759/mda.2022120103 (In Russ., abstr. in Engl.).

References

1. *Riz R*. Natural language processing in Java. DMK-Press. 2016. 264 p.
2. Perspective API. URL: https://www.perspectiveapi.com
3. *van Aken B., Risch J., Krestel R., Löser A.* Challenges for toxic comment classification: An in-depth error analysis. 2018, arXiv:1809.07572.
4. *Risch J., Krestel R.* Toxic Comment Detection in Online Discussions. Deep Learning-Based Approaches for Sentiment Analysis. Springer, Singapore, 2020. P. 85–109.
5. *Weiss K., Khoshgoftaar T.M., Wang D.* A survey of transfer learning // Big Data, 3: 9. 2016. https://doi.org/10.1186/s40537–016–0043–6
6. *Andrusyak B., Rimel M., Kern R.* Detection of Abusive Speech for Mixed Sociolects of Russian and Ukrainian Languages // RASLAN. – 2018. – P. 77–84.

***Evgeniy N. Platonov,** PhD (Physics and Mathematics), Assistant Professor, Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: https://orcid.org/0000-0001-8502-1350, e-mail: en.platonov@gmail.com

****Veronika Y. Rudenko,** Student of the Institute of Information Technologies and Applied Mathematics, Moscow Aviation Institute (National Research University), Moscow, Russia, ORCID: https://orcid.org/0000-0003-0010-331X, e-mail: super.ruden2011@mail.ru

7. *Li Y., Yang T.* Word Embedding for Understanding Natural Language: A Survey. In: Srinivasan S. (eds) Guide to Big Data Applications. Studies in Big Data, vol 26. Springer, Cham. https://doi.org/10.1007/978–3-319–53817–4_4
8. *Liu C.* et al. Research of text classification based on improved TF-IDF algorithm // IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE). 2018 P. 218–222.
9. word2vec // URL: https://code.google.com/archive/p/word2vec/
10. *Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean.* Efficient Estimation of Word Representations in Vector Space // Proceedings of Workshop at ICLR, 2013.
11. *Bojanowski P*, et al. Enriching word vectors with subword information // Transactions of the Association for Computational Linguistics. 2017. V. 5. P. 135–146.
12. *Pennington J., Socher R., Manning C.D. G*love: Global vectors for word representation // Proceedings of the conference on empirical methods in natural language processing (EMNLP). 2014. P. 1532–1543.
13. *Wieting J.* et al. From paraphrase database to compositional paraphrase model and back // Transactions of the Association for Computational Linguistics. 2015. V. 3. P. 345–358.
14. *Chen T., Guestrin C.* Xgboost: A scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. P. 785–794.
15. *Dorogush A.V., Ershov V., Gulin A.* CatBoost: gradient boosting with categorical features support // arXiv preprint arXiv:1810.11363. 2018.
16. *Sepp Hochreiter and Jürgen Schmidhuber.* Long Short-Term memory // Neural computation, V. 9(8). P. 1735–1780, 1997.
17. *Staudemeyer R.C., Morris E.R.* Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks // arXiv preprint arXiv:1909.09586. 2019. URL:https://arxiv.org/pdf/1909.09586.pdf
18. Understanding LSTM Networks URL:https://colah.github.io/posts/2015–08-Understanding-LSTMs/
19. Convolutional Neural Network. An Introduction to Convolutional Neural Networks. URL: https://towardsdatascience.com/convolutional-neural-network-17fb77e76c05
20. *Bai S., Kolter J.Z., Koltun V.* An empirical evaluation of generic convolutional and recurrent networks for sequence modeling // CoRR, abs/1803.01271. 2018. http://arxiv.org/abs/1803.01271
21. Quora Insincere Questions Classification. URL: https://www.kaggle.com/c/quora-insincere-questions-classification/data
22. *T. Fawcett.* An introduction to ROC analysis // Pattern Recognition Letters, V. 27. 2006. P. 861–874.
23. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning. Springer-Verlag, New York. 2017.
24. Eli5 Documentation. URL: https://eli5.readthedocs.io/en/latest/
25. *Tulio Ribeiro M., Singh S., Guestrin C.* Why Should I Trust You? Explaining the Predictions of Any Classifier // KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. P. 1135–1144.
26. glove.840B.300d – pre-trained word vectors GloVe. URL: https://nlp.stanford.edu/projects/glove/
27. wiki-news-300d-1M – pre-trained word vectors trained using fastText. URL: https://fasttext.cc/docs/en/english-vectors.html
28. paragram_300_sl999 – New Paragram-SL999 300 dimensional embeddings tuned on SimLex999 dataset. URL: https://www.kaggle.com/ranik40/paragram-300-sl999
29. GoogleNews-vectors-negative300 – pre-trained word vectors trained using Word2Vec. URL: https://code.google.com/archive/p/word2vec/

Получена 18.01.2022
Принята в печать 31.01.2022

Received 18.01.2022
Accepted 31.01.2022

Моделирование и анализ данных
2022. Том 12. № 1. С. 49–59
DOI: https://doi.org/10.17759/mda.2022120104
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2022. Vol. 12, no. 1, pp. 49–59
DOI: https://doi.org/10.17759/mda.2022120104
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

УДК 004.82

Автоматическое установление родовидовых отношений между понятиями

Макарова А.Е.*
АО «НПК “ВТ и СС”», г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-9232-6838
e-mail: anna20497@list.ru

Никитин Ю.В.**
Институт проблем информатики ФИЦ ИУ РАН
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-7641-0247
e-mail: yuri.v.nikitin@gmail.com

Хорошилов А.А.****
АО «НПК “ВТ и СС”», г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0003-4885-3232
e-mail: a.a.horoshilov@mail.ru

В работе описываются результаты исследований по созданию методов построения формализованного смыслового описания документов для решения задач семантического поиска документов. Применяемые в исследовании методы базируются на использовании процедур машинной грамматики и концептуального анализа текстов, обеспечивающих выявление его понятийного состава и назначения наименованиям понятий характеристик, соответствующих их семантической роли и значимости в тексте. Для выполнения данной работы был создан комплекс программных средств, который был опробован на документах СМИ.

Ключевые слова: семантико-синтаксический анализ, морфологический анализ, концептуальный анализ, автоматическое установление родовидовых отношений.

Для цитаты:
Макарова А.Е., Никитин Ю.В., Хорошилов А.А. Автоматическое установление родовитых отношений между понятиями // Моделирование и анализ данных. 2022. Том 12. № 1. С. 49–59. DOI: https://doi.org/10.17759/mda.2022120104

***Макарова Анна Евгеньевна,** младший программист, АО «НПК “ВТ и СС”», г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-9232-6838, e-mail: anna20497@list.ru

****Никитин Юрий Викторович,** научный сотрудник, Институт проблем информатики ФИЦ ИУ РАН, г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-7641-0247, e-mail: yuri.v.nikitin@gmail.com

*****Хорошилов Александр Александрович,** ведущий программист, АО «НПК “ВТ и СС”», г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0003-4885-3232, e-mail: a.a.horoshilov@mail.ru

1. ВВЕДЕНИЕ

В настоящее время все большую популярность завоевывают технологии семантического полнотекстового поиска, базирующиеся на смысловом анализе информации, в процессе которого производится извлечение смысла из текстовой информации и трансформация его в формальную смысловую модель. Наиболее эффективным лингвистическим инструментом, позволяющим реализовать такую модель поиска, являются так называемые «онтологии» – семантические инструменты, реализующие концептуальные знания о мире в целом и о предметных областях в частности. Описание наиболее общих знаний об окружающем мире содержится в онтологиях верхнего уровня. Знания о конкретных предметных областях содержатся в предметных (тематических) онтологиях.

Следует отметить, что создание онтологических ресурсов требует больших временных затрат и, как показывает практика, их объемы редко превышают несколько десятков тысяч словарных статей. Между тем понятийная система высокотехнологичных отраслей может достигать нескольких сотен тысяч терминологических понятий. Кроме того, в онтологиях наименования понятий представлены в обобщенной форме, а в научно-технических текстах они встречаются в их всевозможных конкретных представлениях, которые не всегда присутствуют в составе конкретной онтологии. Выходом из создавшейся ситуации может быть только автоматизированное создание тематических онтологий большого объема по реальным текстам, относящимся к данной тематике, включая тексты нормативно-технической, проектной и эксплуатационной документации тематических областей.

Построение онтологий – сложный и достаточно длительный процесс, при котором значительные трудозатраты приходятся на человека-эксперта. Чтобы облегчить его труд, в середине 90-х годов начали создаваться программные среды для разработки онтологий. В состав этих сред были включены интерфейсы, которые позволили выполнять ряд операций построения онтологий: концептуализацию, реализацию, проверку непротиворечивости и документирование. За последние годы число инструментов онтологий резко возросло (сайт консорциума W3C, например, предоставляет список более чем 50 инструментов редактирования).

Обзор методов автоматизированного построения онтологий

Существующие методы автоматического построения онтологий можно условно разделить на несколько групп в зависимости от использования основного подхода:

– методы, основанные на лексико-синтаксических шаблонах;
– методы, основанные на системе продукций;
– статистические методы;
– методы, использующие лингвистические подходы.

Подход на основе лексико-синтаксических шаблонов. Данный подход относится к группе методов автоматического построения онтологий, использующих лингвистические средства [1]. Сторонники подхода утверждают, что для построения онтологий следует активно использовать все уровни анализа естественного языка: морфологию, синтаксис и семантику. Подход, основанный на лексико-синтаксических ша-

блонах, давно используется в компьютерной лингвистике. Лексико-синтаксические шаблоны представляют собой характерные выражения и конструкции определенных элементов языка. Данная методика семантического анализа не является ориентированной на определенную предметную область. На основе лексико-синтаксических шаблонов выделяются онтологические конструкции. Ввиду сложности задачи, оценка результатов применения этого подхода проводится опосредованно через анализ результатов его использования, например, в различных приложениях Semantic Web. В целом, лексико-синтаксические шаблоны, как метод семантического анализа данных на естественном языке (в случае большого объема коллекции шаблонов), являются эффективным средством для автоматического построения онтологий.

Подход на основе системы продукций. Данный подход относится к группе методов автоматического построения онтологий, в основе которых лежат подходы из области искусственного интеллекта [2]. Предполагается, что эффективное автоматическое построение онтологий может быть основано на способности методов искусственного интеллекта к извлечению из данных элементов знаний и их нетривиальной переработке. Анализ области естественно-языковой обработки информации показывает преобладание использования различных правил при решении задач в рассматриваемой предметной области. Для создания методов автоматического построения онтологий, как правило, разрабатывается модель генерации системы продукций (на основе применения генетического программирования), модель генерации преобразователей (на основе генетического и автоматного программирования), модель генерации систем логического вывода (также на основе генетического и автоматного программирования) и модель аппарата активации продукций (на основе применения автоматного программирования).

Подход на основе статистических методов. Этот подход к решению проблемы автоматического построения онтологий базируется преимущественно на статистических методах анализа данных на естественном языке [3]. Для его реализации требуются большие объемы текстовой информации (репрезентативные корпуса текстов). При создании такого корпуса текстов требуется его обязательная предварительная обработка. Процесс такой обработки может быть достаточно трудоемким и обычно состоит из нескольких этапов:

– приведение документов к единому формату;
– токенизация;
– стемминг (лемматизация);
– исключение стоп-слов.
– ручная обработка (с использованием инструментов консорциума W3C) по установлению смысловых связей между терминологическими наименованиями понятий.

Однако не всегда есть необходимость в проведении всех вышеперечисленных этапов. В результате предварительной обработки каждый документ коллекции характеризуется вектором типов данного документа и их частотой встречаемости. На первом этапе построения онтологии выделяются входящие в ее состав классы, которые, как правило, базируются на терминах предметной области.

Таким образом, основная задача построения онтологий сводится к выявлению терминов рассматриваемой предметной области. Алгоритмы извлечения терминов из

текстов на естественном языке можно разделить на две группы: статистические и лингвистические. Однако первые обладают определенным преимуществом, поскольку их использование не зависит от лингвистических особенностей конкретного языка. Существующие статистические методы могут показать лучшие результаты, если дополнить их определенными эвристиками.

В качестве базовых эвристик предлагается использовать следующие:

Эвристика № 1.

Имя класса содержит хотя бы одно существительное.

Эвристика № 2.

Общеупотребительные слова по сравнению с терминами обладают большей частотой встречаемости, приблизительно равной в различных предметных областях.

Эвристика № 3.

Считается, что количество информации термина из нескольких слов больше, чем количество информации отдельных слов, входящих в его состав.

Статистический подход базируется на определенной технологической схеме. На первом этапе в каждой коллекции документов выделяют существительные, и определяют их частоту встречаемости. При этом использование частотных критериев значительно сокращает число предполагаемых классов понятий. На втором этапе выделяют термины, состоящие из одного слова. На основании выдвинутой эвристики № 1 сравниваются частоты встречаемости различных существительных в рамках одной коллекции, а также проводится оценка пересечения различных коллекций по используемым существительным (эвристика № 2). Однако статистические данные – не единственный источник классов онтологии.

Терминологические словари также могут стать источниками знаний при формировании ядра онтологии. В случае работы с коллекциями неспециализированных в конкретной области документов возможно использование существующих разработанных экспертами онтологий (например, для английского языка – онтологии WordNet). Наконец, на третьем этапе на основе взаимной информации могут быть выделены термины, состоящие из нескольких слов. Стоит отметить, что в данном случае используется эвристика № 3. Для случая двухсложных терминов получаем, что взаимная информация определяется по формуле:

$$mi(x,y) = P(x,y) / P(x)P(y), \quad (1)$$

где x и y представляют собой отдельные слова термина,

P(x) – частота встречаемости x,

P(y) – частота встречаемости y,

P(x,y) – частота совместной встречаемости x и y.

Выделенные описанным выше образом термины будут представлять собой классы будущей онтологии [2, 3].

Автоматическое формирование тематических словарей на основе синтагматических шаблонов синтагм.

Существенным недостатком всех выше рассмотренных методов создания онтологий требуют значительных трудозатрат экспертов высокой квалификации на раз-

личных этапах этого процесса, что не позволяет построить в приемлемые сроки онтологические ресурсы большого объема. Решить эту проблему возможно только с привлечением достаточно мощных средств автоматизации этого процесса. К таким средствам можно отнести набор базовых средств смыслового анализа текстов: морфологический, семантико-синтаксический и концептуальный анализ текстов.

Авторы настоящего исследования располагали такими средствами [4, 5].

В качества основного инструмента авторами использовался морфологический анализатор МетаФраз [6], семантико-синтаксический анализатор МетаФраз [7] и разработанная авторами настоящей статьи процедура концептуального анализа текстов, базирующаяся на синтагматических шаблонах. Положенные в основу метода этой процедуры синтагматические шаблоны функционируют на основе метода лингвистической аналогии текстовых форм представлений наименований понятий в текстах и по сути представляют собой синтаксическую модель фрагмента текста в рамках следующего утверждения: представление синтаксической структуры фрагментов текстов в виде последовательности контактно расположенных двухбайтовых индексов элементов синтагм, обладающих грамматическими свойствами конкретных слов-эталонов, позволяет фиксировать грамматические и синтаксические свойства различных последовательностей реальных текстов, а также позволяет в ряде задач распознавать аналогичные по заданным свойствам последовательности слов и словосочетаний в текстах [7].

В качестве эталонных словосочетаний были использованы элементы эталонного концептуального словаря наименований понятий (ЭКСНП) объемом 1.7 млн. словарных статей. В таблице 1 приведен фрагмент понятийной системы текста по медицинской тематике, выявленный на основе обобщенных синтагм, полученных по словарю ЭКСНП.

Таблица 1

**Фрагмент понятийной системы текста,
выявленный на основе обобщенных синтагм**

HAAwHAAwABNw	острого респираторного синдрома / атомного ракетного крейсера
HAAwABNw	респираторного синдрома / абсолютного принципа
HIAwwмNw	атипичная пневмония / абдоминальная аортография
HTAwAANw	воспалительный процесс / абразивный износ
HTAwAANwHAAwABNw	воспалительный процесс пищеварительного тракта / преступный акт пещерного садизма
AANwHAAwABNw	процесс пищеварительного тракта / анализ клеточного цикла
HAAwABNw	пищеварительного тракта / абсолютного принципа
HBAwwfNw	коронавирусной инфекции / аварийной ситуации
HIAwHIAw	острая респираторная / вегетативная нервная

Словарная запись в таблице 1 каждого элемента состоит их трех компонент: шаблона синтаксической синтагмы словосочетания (первая слева запись), далее в центре выделенное из текста словосочетание, и последнее справа словосочетание (через косую черту) – словосочетание эталонного словаря.

Автоматизированное составление словарей терминологических наименований понятий по текстам документов можно выполнить по следующей технологической схеме:

1. Формально-логический контроль текстов;
2. Членение исходного текста на отдельные слова;
3. Морфологический анализ слов текста;
4. Членение текста на предложения;
5. Семантико-синтаксический анализ текстов;
6. Приближенный концептуальный анализ текстов;
8. Автоматическое приведение наименований понятий к их канонической форме;
9. Формирование частотного словаря наименований понятий.

2. АЛГОРИТМ УСТАНОВЛЕНИЯ РОДОВИДОВЫХ ОТНОШЕНИЙ В ТЕКСТАХ

Полученный понятийный словарь слов и словосочетаний текста достаточно большого объема или совокупной тематической коллекции текстов дает возможность автоматически устанавливать смысловые отношения между словосочетаниями на основе анализа лексического состава словосочетаний. Для установления родовидовых отношений между словосочетаниями можно воспользоваться следующей гипотезой: если имеются два словосочетания различной длины, в которых имеются одинаковые главные (опорные) слова, и все определители более короткого словосочетания совпадают с определителями более длинного словосочетания, то эти два словосочетания находятся вотношении родовидовой связи, и более короткое словосочетание является родовым понятием, а более длинное – видовым;

В соответствие с этой гипотезой был разработан следующий алгоритм установления родовидовых отношений в текстах между наименованиями понятий:

Устанавливаем последовательность слов и словосочетаний, имеющих одно и то же главное слово. Обрабатываемое словосочетание располагается справа от разделителей звездочек, а его главное слово слева. Результаты занесем в таблицу 2.

Таблица 2

Список словосочетаний, использованных для исследования, и их главное слово

академия***Академии наук
академия***Академия Генштаба
академия***Киево-Могилянская академия
академия***Российская Академия
академия***академия МВД
академия***военная академия
академия***сельскохозяйственная академия
академия***Академия Народного Хозяйства
академия***Академия военных наук
академия***Академия наук Татарстана
академия***Военная академия Генштаба
академия***Польской академии наук
академия***Российская академия народного хозяйства
академия***Украинская свободная академия наук

Строим матрицу M вхождений отдельных слов в словосочетания. Номера строк матрицы i – это индексы предложений (или словосочетаний) $i = 1, 2, \ldots$ n , где n – количество обрабатываемых предложений или словосочетаний. Номера столбцов матрицы j – индексы уникальных слов в тексте $j = 1, 2, \ldots$ k , где k – количество всех уникальных слов.

Для заполнения используются нули и единицы. Если в предложении i , присутствует слово с индексом j , в ячейке a_{ij} ставится единица, в противном случае ноль.

$$M = \begin{pmatrix} 1 & \cdots & a_k \\ \vdots & \ddots & \vdots \\ a_n & \cdots & a_{nk} \end{pmatrix}$$

1. Последовательно устанавливаем вхождения N слов в словосочетания, начиная с N = 1. Первым в обработку поступает слово, встречающееся чаще всего, то есть главное. В нашем примере это «академия».
2. Если все слова словосочетания имеют частоту равную единице, то это будут словосочетания 1-го уровня RV отношений. Родовым понятием для них будет одно это слово (если такого родового слова нет, то его необходимо создать принудительно). Приведем пример словосочетаний, относящихся к первому уровню RV отношений, выбрав их из Таблицы 2:
   - Академии наук
   - Академия Генштаба
   - Киево-Могилянская академия
   - Российская Академия
   - академия МВД
   - военная академия
   - сельскохозяйственная академия
   - Академия Народного Хозяйства
3. Далее в соответствии с матрицей отдельных слов в словосочетаниях устанавливаются отношения следующего уровня RV отношений (N+1), путем выявления вхождения слов в словосочетания, имеющих частоту f > 1.
   Примером словосочетаний с RV отношениями второго уровня можно считать:
   Для родового понятия «академия наук»:
   - Академия военных наук
   - Академия наук Татарстана
   - Польской академии наук
   - Украинская свободная академия наук

   Для родового понятия «академия Генштаба»:
   - Военная академия Генштаба

   Для родового понятия «военная академия»:
   - Академия военных наук
   - Военная академия Генштаба

   Для родового понятия «Российская Академия»:
   - Российская академия народного хозяйства

4. Для каждой цепочки устанавливаем словосочетание с минимальной длиной и полностью входящее в другое словосочетание. Для каждой пары продолжаем устанавливать следующий уровень RV отношений.
   Заметим, что в нашем примере возможно выделить квазиродовое понятие «академия народного хозяйства», входящее в состав словосочетания «Российская академия народного хозяйства». Они связаны RV отношениями третьего уровня.
5. Процесс установления RV отношений заканчивается, когда будут исчерпаны все понятия с частотной лексикой.
   В таблице 3 приведены результаты работы алгоритма.

Таблица 3

Результаты работы алгоритма

академия =RV= Академии наук
академия =RV= академия МВД
академия =RV= Киево-Могилянская академия
академия =RV= сельскохозяйственная академия
академия =RV= Академия Народного Хозяйства
академия =RV= Академия Генштаба
академии наук =RV= Академия военных наук
академии наук =RV= Польской академии наук
академии наук =RV= Украинская свободная академия наук
академии наук =RV= Академия наук Татарстана
академия Генштаба =RV= Военная академия Генштаба
военная академия =RV= Академия военных наук
военная академия =RV= Военная академия Генштаба
российская Академия =RV= Российская академия народного хозяйства
академия Народного Хозяйства =RV= Российская академия народного хозяйства

3. ЗАКЛЮЧЕНИЕ

В проведенном исследовании показана принципиальная возможность на основе предлагаемых методов автоматически создать программные и декларативные средства для процедур автоматического семантического поиска документов, в частности, для процедур автоматического установления степени смысловой близости документов. На основе созданного в процессе проведения настоящего исследования алгоритма автоматического установления родовидовых отношений между понятиями был проведен эксперимент по обработке массива текстов СМИ большого объема. В ходе эксперимента получены удовлетворительные результаты. Их анализ показал, что на количественные характеристики результатов обработки текстов незначительно повлияли ресурсные ограничения и принятые в связи с этим принятые допущения.

Для улучшения работы используемых процедур необходимо предпринять следующие шаги:

1. Необходимо выполнить в полном объеме комплекс технологических операций по созданию декларативных средств. Исходные тексты предварительно должны быть

подвергнуты обработке процедурами формально-логического контроля и исправления орфографических и синтаксических ошибок. Большая часть процедур анализа и исправления текстовых искажений должны быть автоматизированы.

2. Необходимо обеспечить требуемые параметры концептуальных словарей (по объему и составу). При этом необходимо исходить из следующих рекомендаций: тематический словарь должен быть составлен по актуальным текстам и иметь объем не менее 1 млн. словарных статей. Покрытие анализируемых текстов наименованиями понятий должно быть не менее 60–70 % от их общего состава
3. Для повышения степени обобщения смыслового содержания наименований понятий необходимо, чтобы наименования понятий в словарях были связаны также отношениями синонимии.

Литература

1. *Захарова И.В.* Математическая модель семантического поиска с использованием онтологического подхода: Автореф. дис. канд. физ.-мат. наук. – Челябинск, 2009. – 20 с.
2. *Найханова Л.В.* Методы и модели автоматического построения онтологий на основе генетического и автоматного программирования: Автореф. дис. докт. тех. наук. – Красноярск, 2008. – 36 с.
3. *Рабчевский Е.А.* Автоматическое построение онтологий. // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. – СПб.: Издательство Политехнического Университета, 2007. – № 52–2. – С. 22–26.
4. *Хорошилов А.А.* Методы автоматического установления смысловой близости документов на основе их концептуального анализа // Труды XV Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, г. Ярославль, 14–17 октября 2013 года. – 2013. – С. 369–376.
5. *Хорошилов А.А., Макарова А.Е.* Автоматизированное формирование онтологических ресурсов в авиакосмической отрасли // Тезисы. – М.: Издательство «Перо», 2021– 9,43 Мб. [Электронное издание]. – Москва: Московский авиационный институт (национальный исследовательский университет), 2021. – С.282–283.
6. Морфологический анализатор МетаФраз нового поколения / Хорошилов Ал-др А., Никитин Ю.В., Пшеничный С.И., Шевкунов М.А., Хорошилов Ал-ей А. // Научно-техническая информация. Серия 2: Информационные процессы и системы. ВИНИТИ РАН. 2021. № 5.
7. Автоматическое формирование синтаксической модели языка для задач машинного перевода и информационного поиска / Кан А.В., Ревина В.Д., Руснак В.И., Хорошилов Ал-др А., Хорошилов Ал-сей А. // Сб. «Научно-техническая информация», Сер. 2, № 12, ВИНИТИ, 2018.

Automatic Establishment of Generic Relationships Between Concepts

Anna E. Makarova*
AO ″NPK “VT i SS”″, Moscow, Russia
ORCID: https://orcid.org/0000-0002-9232-6838
e-mail: anna20497@list.ru

Yuriy V. Nikitin**
Institute for Informatics Problems FITS IU RAN, Moscow, Russia
ORCID: https://orcid.org/0000-0002-7641-0247
e-mail: yuri.v.nikitin@gmail.com

Alexsander A. Khoroshilov****
AO ″NPK “VT i SS”″, Moscow, Russia
ORCID: https://orcid.org/0000-0003-4885-3232
e-mail: a.a.horoshilov@mail.ru

The article describes the results of research on the creation of methods for constructing a formalized semantic description of documents for solving problems of semantic search for documents. The methods used in the study are based on the use of machine grammar procedures and conceptual analysis of texts, which ensure the identification of its conceptual composition and the assignment of characteristics to the names of concepts that correspond to their semantic role and significance in the text. To perform this work, a set of software tools was created, which was tested on media documents.

Keywords: semantic-syntactic analysis, morphological analysis, conceptual analysis, automatic establishment of generic relations.

For citation:
Makarova A.E., Nikitin Yu.V., Khoroshilov A.A. Automatic Establishment of Generic Relationships Between Concepts. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*, 2022. Vol. 12, no. 1, pp. 49–59. DOI: https://doi.org/10.17759/mda.2022120104 (In Russ., abstr. in Engl.).

References

1. Zakharova I.V. Matematicheskaya model’ semanticheskogo poiska s ispol’zovaniem ontologicheskogo podkhoda: Avtoref. dis. kand. fiz.-mat. nauk. – Chelyabinsk, 2009. – 20 p.
2. Naikhanova L.V. Metody i modeli avtomaticheskogo postroeniya ontologii na osnove geneticheskogo i avtomatnogo programmirovaniya: Avtoref. dis. dokt. tekh. nauk. – Krasnoyarsk, 2008. – 36 p.

***Anna E. Makarova,** Junior programmer, AO ″NPK “VT i SS”″, Moscow, Russia, ORCID: https://orcid.org/0000-0002-9232-6838, e-mail: anna20497@list.ru

****Yuriy V. Nikitin,** Researcher, Institute for Informatics Problems FITS IU RAN, Moscow, Russia, ORCID: https://orcid.org/0000-0002-7641-0247, e-mail: yuri.v.nikitin@gmail.com

*****Alexsander A. Khoroshilov,** Senior programmer, AO ″NPK “VT i SS”″, Moscow, Russia, ORCID: https://orcid.org/0000-0003-4885-3232, e-mail: a.a.horoshilov@mail.ru

3. Rabchevskii E.A. Avtomaticheskoe postroenie ontologii. Nauchno-tekhnicheskie vedomosti Sankt-Peterburgskogo gosudarstvennogo politekhnicheskogo universiteta. – SPb.: Izdatel'stvo Politekhnicheskogo Universiteta, 2007. – № 52–2. – P. 22–26.
4. Khoroshilov A.A. Metody avtomaticheskogo ustanovleniya smyslovoi blizosti dokumentov na osnove ikh kontseptual'nogo analiza. Trudy XV Vserossiiskoi nauchnoi konferentsii «Ehlektronnye biblioteki: perspektivnye metody i tekhnologii, ehlektronnye kollektsiI» – RCDL'2013, g. Yaroslavl', 14–17 oktyabrya 2013 goda. – 2013. – P. 369–376.
5. Khoroshilov A.A., Makarova A.E. Avtomatizirovannoe formirovanie ontologicheskikh resursov v aviakosmicheskoi otrasli. Tezisy. – M.: Izdatel'stvo «PerO», 2021. – 9,43 Mb. [Ehlektronnoe izdanie]. – Moskva: Moskovskii aviatsionnyi institut (natsional'nyi issledovatel'skii universitet), 2021. – P. 282–283.
6. Morfologicheskii analizator MetAFraz novogo pokoleniya / Khoroshilov Al-dr A., Nikitin YU.V., Pshenichnyi S.I., Shevkunov M.A., Khoroshilov Al-ei A. Nauchno-tekhnicheskaya informatsiya. Seriya 2: Informatsionnye protsessy i sistemy. VINITI RAN. 2021. № 5.
7. Avtomaticheskoe formirovanie sintaksicheskoi modeli yazyka dlya zadach mashinnogo perevoda i informatsionnogo poiska. Kan A.V., Revina V.D., Rusnak V.I., Khoroshilov Al-dr A., Khoroshilov Al-sei A. Sb. «Nauchno-tekhnicheskaya informatsiYA», Ser. 2, № 12, VINITI, 2018.

Получена 18.02.2022
Принята в печать 14.03.2022

Received 18.02.2022
Accepted 14.03.2022

Моделирование и анализ данных
2022. Том 12. № 1. С. 60–78
DOI: https://doi.org/10.17759/mda.2022120105
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2022. Vol. 12, no. 1, pp. 60–78
DOI: https://doi.org/10.17759/mda.2022120105
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

◇◇◇◇◇◇◇◇МЕТОДИКА ОБУЧЕНИЯ◇◇◇◇◇◇◇◇

УДК 371.3

Особенности обучения студентов с ОВЗ по зрению дисциплинам математического и компьютерного циклов на факультете «Информационные технологии» с применением дистанционных технологий

Червен-Водали Е.Б.*
Московский государственный психолого-педагогический университет (ФГБОУ ВО МГППУ), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-6871-9105
e-mail: cervenvodali@mgppu.ru

Антипова С.Н.**
Московский государственный психолого-педагогический университет (ФГБОУ ВО МГППУ), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0001-6642-7953
e-mail: antipovasn@mgppu.ru

Сидорова В.Б.***
Московский государственный психолого-педагогический университет (ФГБОУ ВО МГППУ), г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0001-6391-5361
e-mail: sidorovavb@mgppu.ru

В статье рассматриваются вопросы организации образовательного процесса в условиях дистанционного обучения для студентов-инвалидов и лиц с ОВЗ по зрению в связи с введением режима самоизоляции в период заболеваемости Covid-19. Особый акцент делается на преподавании дисциплин компьютерного и математического цикла студентам с нарушением зрения, так как эти дисциплины имеют свою специфику.

Ключевые слова: образовательный процесс, дистанционные технологии, студенты с ОВЗ.

Для цитаты:
Червен-Водали Е.Б., Антипова С.Н., Сидорова В.Б. Особенности обучения студентов с ОВЗ по зрению дисциплинам математического и компьютерного циклов на факультете «Информационные технологии» с применением дистанционных технологий // Моделирование и анализ данных. 2022. Том 12. № 1. С. 60–78. DOI: https://doi.org/10.17759/mda.2022120105

***Червен-Водали Елена Борисовна,** преподаватель кафедры прикладной информатики и мультимедийных технологий, Московский государственный психолого-педагогический университет (ФГБОУ ВО МГППУ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-6871-9105, cervenvodali@mgppu.ru

****Антипова Светлана Николаевна,** заместитель декана по внеучебной работе факультета информационных технологий, Московский государственный психолого-педагогический университет, г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0001-6642-7953, e-mail: antipovasn@mgppu.ru

*****Сидорова Валерия Борисовна,** преподаватель кафедры прикладной информатики и мультимедийных технологий, Московский государственный психолого-педагогический университет (ФГБОУ ВО МГППУ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0001-6391-5361 , e-mail: sidorovavb@mgppu.ru

1. ВВЕДЕНИЕ

С появлением новых информационно-коммуникативных технологий появилась необходимость введения в традиционный образовательный процесс новых методов и средств обучения. Очное присутствие в аудитории больше не является единственным вариантом обучения. Внедрение цифровых технологий в образовательный процесс – одно из приоритетных направлений развития современной системы образования. Из-за введения режима самоизоляции все эти процессы заметно ускорились. Пополнить свои знания, узнать что-то новое можно в любое время и в любом месте, если есть доступ к любому электронному устройству с хорошим выходом в интернет. Некоторые проводимые исследования по использованию мобильных технологий в онлайн режиме показали, что интерес к учебе у студентов повышается, а, с другой стороны, живую коммуникацию никакое удаленное обучение заменить не сможет. Особенно элементы дистанционного образования идеально подходят тем людям, которые уже работают или имеют другие личные обстоятельства, по которым не имеют возможности присутствовать на занятиях очно, к такой категории людей относятся и люди, имеющие ограничения по состоянию здоровья. Цифровые технологии с каждым годом открывают все больше новых возможностей. И даже тем, кто был далек от этого направления, приходиться постигать новые вершины информационного образования. Каждый участник современного образовательного процесса выбирает для себя наиболее удобные, понятные и продуктивные технологии.

В связи с введением режима самоизоляции в период заболеваемости Covid-19 факультет Информационных технологий, как и большинство ВУЗов страны, перешел на дистанционный формат обучения. Так как все были не готовы к такой ситуации, то все инструменты подачи материала в онлайн режиме подбирались методом проб и ошибок. Особый акцент делался на преподавании дисциплин компьютерного и математического цикла студентам с нарушением зрения, так как эти дисциплины имеют свою специфику: формулы, графики, схемы и т.п.

Будем рассматривать дистанционное или онлайн-обучение как электронное обучение, которое проводится посредством дистанционных образовательных технологий.

Закон «Об образовании» четко определяет понятия «электронное обучение» и «дистанционные образовательные технологии»:

Статья 16. Реализация образовательных программ с применением электронного обучения и дистанционных образовательных технологий.

«Под электронным обучением понимается организация образовательной деятельности с применением содержащейся в базах данных и используемой при реализации образовательных программ информации и обеспечивающих ее обработку информационных технологий, технических средств, а также информационно-телекоммуникационных сетей, обеспечивающих передачу по линиям связи указанной информации, взаимодействие обучающихся и педагогических работников.

Под дистанционными образовательными технологиями понимаются образовательные технологии, реализуемые в основном с применением информационно-телекоммуникационных сетей при опосредованном (на расстоянии) взаимодействии обучающихся и педагогических работников».

Следует отметить, что дистанционное обучение и электронное обучение можно рассматривать как отдельные виды обучения, но та модель обучения, с которой пришлось столкнуться, предполагает их тесную взаимосвязь. В статье говорится об организации образовательного процесса в условиях дистанционного обучения для студентов-инвалидов и лиц с ОВЗ по зрению на факультете «Информационные технологии». Дистанционная поддержка учебных дисциплин позволяет построить эффективную модель инклюзивного образования, на основе интеграции электронного и традиционного обучения. Дистанционное обучение подразумевает целенаправленную работу, которая направлена для обеспечения комфортного обучения студентам с ограниченными возможностями здоровья. Внедрение дистанционного обучения и новейших технологий в систему образования позволяет приобретать студентам нужные навыки и компетенции также как и при очном образовании, а также это подразумевает возможность студенту с ОВЗ научиться использовать информационные ресурсы сети Интернет в профессиональной деятельности, а в дальнейшем осуществлять поиск, проводить анализ и оценку информации, сформировать навыки профессионального общения. Организация образовательного процесса по программам высшего образования для лиц с ограниченными возможностями здоровья направлены на создание условий, обеспечивающих получения ими профессиональной подготовки и профессионального образования с учетом требований рынка труда и перспектив развития профессий, а также условий для их социальной адаптации и интеграции в общественную инфраструктуру.

Для обеспечения доступности и качества образования большое значение имеет применение информационных и коммуникационных технологий (ИКТ).

В области применения ИКТ в образовании людей с особыми потребностями чрезвычайно разнообразны, можно выделить следующие основные направления в их использовании:

1. **Применение ИКТ для решения компенсаторных задач.** Использование технологий в качестве вспомогательных устройств позволяет учащимся с особыми

потребностями принимать активное участие в учебном процессе и коммуникации. Так, например, в случае двигательных нарушений ИКТ позволяют человеку писать, а в случае зрительных нарушений делают возможным процесс чтения. С этой точки зрения технологии обеспечивают учащимся возможность контролировать окружающую среду, позволяют решать учебные и социальные задачи, предоставляют доступ к информационным ресурсам.

2. **Применение ИКТ для решения дидактических задач.** Использование ИКТ как дидактического инструмента способствует изменению подходов к учебно-методическому процессу и стимулирует появление новых стратегий обучения и контроля знаний. Использование информационных технологий позволяет свести к минимуму различия между учащимися и делает возможным применение современных педагогических приемов, направленных на улучшение взаимодействия учащихся между собой и с преподавателями.
3. **Применение ИКТ для решения коммуникативных задач.** Технологии могут выступать посредниками в процессе общения людей с особыми потребностями. Для каждой категории пользователей, испытывающих трудности в процессе коммуникации, должны быть подобраны и адаптированы специальные вспомогательные устройства и программное обеспечение. Прежде всего, речь идет о компьютере для людей, у которых трудности коммуникации осложняются зрительными нарушениями, технологии зачастую являются единственным способом связи с внешним миром, позволяющим выразить свои мысли и потребности. Основными типами средств ИКТ, используемыми для обучения инвалидов, являются следующие:
   - стандартные технологии – например, компьютеры, имеющие встроенные функции настройки для лиц с ограниченными возможностями здоровья;
   - доступные форматы данных, известные также как альтернативные форматы – например, доступный HTML, говорящие книги системы DAISY (Digital Accessibility Information System – электронная доступная информационная система); а также «низкотехнологичные» форматы, такие как система Брайля;
   - вспомогательные технологии: устройства для чтения с экрана, клавиатуры со специальными возможностями и т.д. Вспомогательные технологии (ВТ) – это «устройства, продукты, оборудование, программное обеспечение или услуги, направленные на усиление, поддержку или улучшение функциональных возможностей людей с ограниченными возможностями здоровья».

Одним из требований Федерального образовательного стандарта высшего образования (ФГОС ВО) является то, что обучающиеся из лиц с ограниченными возможностями здоровья должны быть обеспечены печатными и (или) электронными образовательными ресурсами в формах, адаптированных к ограничениям из здоровья.

Компьютерные тифлотехнологии базируются на комплексе аппаратных и программных средств, обеспечивающих преобразование компьютерной информации в доступные для незрячих и слабовидящих формы (звуковое воспроизведение, рельефно-точечный или укрупненный текст), и позволяют им самостоятельно работать на обычном персональном компьютере с программами общего назначения, Тифло-

технические средства, используемые в учебном процессе студентов с нарушениями зрения, условно делятся на две группы: средства для усиления остаточного зрения и средства преобразования визуальной информации в аудио и тактильные сигналы. Для слабовидящих студентов в лекционных и учебных аудиториях необходимо предусмотреть возможность просмотра удаленных объектов (например, текста на доске или слайда на экране) при помощи видеоувеличителей для удаленного просмотра.

Изучение математических дисциплин для студентов с нарушением зрения сопряжено со значительными трудностями. При обучении дисциплинам из области математики основными являются визуальные источники информации – записи формул на доске, плоскопечатные учебники. Не имея возможности следить за записью преподавателя, незрячие студенты вынуждены воспринимать лекционный материал на слух. Конспектирование материала студенты ведут одновременно двумя доступными им способами: запись услышанного с помощью письменных принадлежностей по системе Брайля и ведение аудиозаписи происходящего в аудитории. Т.е. незрячие студенты полностью полагаются на речь преподавателя при получении лекционного материала. Громоздкость письменных принадлежностей и специфика записи по системе Брайля не позволяют вести хороших конспективных материалов. Это можно назвать рабочим черновиком, слабо способным помочь студенту при подготовке к экзамену или при выполнении домашнего задания. Следует отметить, что многие лекторы вообще не произносят все формулы, написанные на доске. В этом случае незрячий вообще лишен возможности записать материал лекции.

Темп письма по рельефно-точечной системе Брайля ниже, чем темп письма обычной авторучкой. Зрячие студенты имеют возможность проработать пройденный материал по учебникам, которые любая ВУЗовская библиотека имеет в достаточном количестве. Незрячие студенты лишены такой возможности. Хотя и существует учебная литература по высшей математике для незрячих, но это всего лишь несколько книг, изданных еще в СССР. В настоящее время в связи с высокой себестоимостью и трудоемкостью издания, математическая литература рельефно-точечным шрифтом не издается. Немногие, сохранившиеся до настоящего времени учебники можно взять только в специальных библиотеках для незрячих, ни один ВУЗ не может предоставить студентам с нарушением зрения учебную литературу в доступной форме.

Необходимо сделать информационную среду более доступной для студентов с нарушением зрения. Тем самым, будет внесен значительный вклад в расширение возможностей людей, имеющих проблемы со зрением. В дальнейшем, открытие новых возможностей для инвалидов отзовется новыми научными работами при участии этой категории граждан. Заинтересованность в улучшении условий и качестве обучения студентов с нарушением зрения способствует созданию современных методов адаптации текстовой и графической информации. Необходимо привлечь все современные технические средства для создания доступной информационной среды, которая является основой системы высшего образования.

Доступные формы представления информации

Чтобы быть доступным для незрячего, материал может быть представлен в двух видах: рельефно-точечные записи шрифтом Брайля или аудиальной речевой

информации. Существуют современные компьютерные технологии, позволяющие сканировать и озвучивать плоскопечатную информацию с помощью специализированного программного обеспечения. Но эти технологии не позволяют озвучить математические формулы, схемы, графики и любую другую не текстовую информацию. Таким образом они совершенно не подходят для студентов, изучающих математические дисциплины.

На факультете разработана технология изготовления учебно-методических комплексов для студентов со зрительными патологиями, слушающих курс математики. При разработке данной технологии были решены следующие задачи:

- Исследование общих принципов записи математических выражений с помощью системы Брайля;
- Оценка существующих программных средств подготовки математических текстов для печати на брайлевском принтере;
- Разработка программного обеспечения для преобразования файлов в формате TEX в файлы для распечатки рельефно-точечным шрифтом Брайля и в файлы для чтения программой речевого доступа Jaws for Windows;
- Использование брайлевского принтера для мелкотиражного производства материала, отпечатанного по системе Брайля;
- Разработка программного обеспечения для навигации по аудио файлам;
- Адаптация и разметка аудиозаписи для работы в режиме учебного пособия.

Таким образом, удалось разработать и внедрить технологию изготовления учебно-методических комплексов для студентов с нарушением зрения. Используя особенности издательской системы TEX данная технология позволяет подготовить в доступной для незрячего форме любые математические тексты. Используются обе доступные формы информации – рельефно-точечная и речевая. Причем преобразование в речевую форму может быть осуществлено как диктором, так и в автоматическом режиме с помощью программного синтезатора речи. Второй способ преобразования в речь, хотя гораздо быстрее и дешевле, но для конечного пользователя менее предпочтителен так, как синтезированная речь менее разборчива и гораздо больше утомляет слушателя.

2. СИСТЕМА LATEX И АНАЛОГИИ С СИСТЕМОЙ БРАЙЛЯ

Tex – это компьютерная программа, созданная Дональдом Кнутом, предназначенная для верстки текста и математических формул. Кнут начал писать Tex в 1977 году. Tex, в том виде, в каком он используется сегодня, был выпущен в 1982 году и слегка улучшен с годами. Последние несколько лет Tex стал чрезвычайно стабилен. Кнут утверждает, что в нем практически нет ошибок.

LaTex – макропакет, позволяющий авторам верстать в печать их работы с высоким типографским качеством, при помощи заранее определенных, профессиональных макетов. LaTex подходит так же для создания книг содержащих большое коли-

чество математических формул и выражений. Первая версия была выпущена Лесли Лампортом в 1984 г. Пакет позволяет автоматизировать многие задачи набора текста и подготовки статей, включая набор текста на нескольких языках, нумерацию разделов и формул, перекрёстные ссылки, размещение иллюстраций и таблиц на странице, ведение библиографии и др. Кроме базового набора существует множество пакетов расширения LaTeX. Текущая версия – LaTeX2ε, после создания в 1994 году испытывала некоторый период нестабильности, окончившийся к концу 90-х годов, а в настоящее время стабилизировалась (хотя раз в год выходит новая версия).

Общий внешний вид документа в LaTeX определяется стилевым файлом. Существует несколько стандартных стилевых файлов для статей, книг, писем и т.д., кроме того, многие издательства и журналы предоставляют свои собственные стилевые файлы, что позволяет быстро оформить публикацию, соответствующую стандартам издания.

Между системой LaTex и системой Брайля есть много общего. Так, например, в системе LaTex есть специальные символы, которые имеют особое значение. Если ввести их в тексте напрямую, то они обычно не печатаются. Можно привести несколько примеров таких символов: «_» – нижний индекс, «^» – верхний индекс. В системе Брайля так же используются символы для обозначения верхних и нижних индексов. Есть и еще одно сходство, позволившее реализовать преобразование в систему Брайля. Информация в формате TEX представляется линейно, что упрощает процесс преобразования TEX-файлов в речь или файлы для печати рельефно-точечным шрифтом Брайля.

Система Брайля стала единственным способом записи текстовой информации на бумагу незрячими людьми. Запись математических формул с помощью системы Брайля – процесс более сложный в отличии от литературного текста. В настоящее время, при письме по системе Брайля сохранились принципы записи, сложившиеся и утверждённые комиссией Всероссийского Общества Слепых ещё в 70-х годах прошлого столетия. В те годы была выпущена книга «советская система обозначений для слепых по математике и другим естественным наукам». Разобраться человеку, не знакомому с принципами письма по системе брайля, в записи математических выражений очень сложно.

В основе записи многострочных и многоуровневых выражений, таких как дробь, интеграл, предел и др., лежит принцип линеаризации.

Письмо системой Брайля идёт по строго разграниченным строкам. Это связано с относительной ориентацией символов в тексте и удобством поиска следующего символа, что делает недопустимым смещение текста относительно оси строки вверх или вниз. Иначе говоря, читатель перемещает пальцы только в направлении слева на право по строке, чтобы перейти к следующему знаку, и, достигнув конца строки, переходит к началу следующей.

Построчная запись всего материала вынуждает незрячих переструктурировать принципы записи многих математических выражений. Тем не менее, упорядочив правила, мы имеем возможность получить аналоги всех математических выражений в рельефно-точечном шрифте Брайля.

Рассмотрим структуру любого математического выражения. Иначе говоря, представим его в виде составных частей. При записи авторучкой мы последовательно вырисовываем символы, составляющие это выражение. Любую математическую запись можно рассматривать как функцию с несколькими параметрами. Например, для задания дроби мы должны определить два параметра: числитель дроби и знаменатель дроби. Если оба параметра указаны, то данное выражение однозначно определено. Теперь указав общий синтаксис и порядок записи параметров для дробей, мы получим правило записи дробей. Можно провести аналогию с чтением дробей, ведь сначала читается числитель дроби, а затем знаменатель. Максимально сохраняя порядок чтения с порядком написания, мы установим очерёдность записи составляющих элементов математических выражений.

Учитывая вышесказанное, приходим к тому, что общий вид записи дроби по системе Брайля имеет следующий вид:

Сначала, записывается спецзнак, предупреждающий о том, что началась запись дроби. Далее в строчку, числитель дроби, затем специальный символ «дробной черты», а затем знаменатель. Запись дроби оканчивается опять же спецзнаком, говорящим о том, что дробь окончилась.

Тем самым, дробь записывается в одну строку, т.е. Линеаризуется.

Это лишь немногое, что можно выделить из множества правил рельефно-точечной системы записи математических выражений. Многоуровневые выражения, такие как дроби, интегралы, выражения с показателем степени и т.п. невозможно записать так, как это можно записать в плоскопечатном виде. Все эти выражения записываются в строчку. Поддерживается только линейная запись.

Используя сходство системы TEX и рельефно-точечной системы Брайля на факультете информационных технологий МГППУ была создана программа TeXToBraille. Программа TeXToBraille (TeX в Брайль) принимает на вход файл, размеченный по правилам языка LaTeX. За основу синтаксиса взят язык, описанный в книге С. Львовского «Набор и верстка в пакете LaTeX». Используя словарь макросов, программа в результате обработки файла подставляет вместо макросов LaTeX описательные конструкции на русском языке, или формирует запись по правилам системы Брайля. В общем виде результат работы программы имеет вид: «Начало выражения ... вторая часть выражения ... третья часть выражения ... конец выражения» Более конкретный пример: выражение «\sqrt{x^2+y^2}» после обработки превращается в выражение: «Корень из х малое в степени 2 +у малое в степени два конец корня».

В случае описательных выражений на русском языке текст может быть прослушан с помощью программного синтезатора речи на персональном компьютере, в случае конвертации исходного файла LaTeX в синтаксис системы Брайля текст может быть распечатан на специальном принтере.

Система может использоваться в учебных учреждениях, в специализированных издательствах и библиотеках, частными лицами для подготовки учебных и научных материалов по разделам высшей математики для печати рельефным шрифтом Брайля с целью последующего использования их незрячими и слабовидящими школьниками, студентами, специалистами в ходе учебной или профессиональной деятельности.

Программа обрабатывает следующие выражения и формулы:

- Интегралы;
- Пределы;
- Дробные выражения;
- Верхние и нижние индексы;
- Степени с различными показателями;
- Корни.

Целесообразно, материалы конспективного плана производить по месту обучения с учётом специфики читаемого курса. Тем самым, будет достигнут максимальный эффект от мелкотиражного производства учебной и методической литературы. Изготовленные таким образом методические пособия позволят оперативно восполнить недостаток литературы, так необходимой незрячим студентам для подготовки к экзаменам.

3. ИЗУЧЕНИЕ МАТЕМАТИЧЕСКИХ ТЕКСТОВ С ПОМОЩЬЮ ПРОГРАММЫ РЕЧЕВОГО ДОСТУПА JAWS

При работе за компьютером у зрячего пользователя основным устройством вывода информации является монитор. Незрячий студент использует специальное программное обеспечение, осуществляющее озвучивание информации на экране компьютера. Эти программы позволяют контролировать информацию, вводимую с клавиатуры и выводимую на экран персонального IBM-совместимого компьютера в текстовом режиме. Это дает незрячему пользователю возможность работы с любыми программами различного назначения, например, с текстовыми и табличными процессорами, системами программирования.

Основные функции программ речевого доступа:

- озвучивание информации, вводимой с клавиатуры;
- автоматическое озвучивание текстовой информации, выводимой на экран другими программами;
- чтение фрагментов экрана по командам пользователя (символа, слова, строки, заданной области и т.д.) в процессе функционирования других прикладных и системных программ;
- отслеживание изменений на экране и оповещение о них пользователя;
- автоматическая загрузка конфигураций, приспособленных для работы с конкретной прикладной программой при ее запуске.

Незрячие пользователи, работающие с операционной системой Windows, используют одну из таких программ «JAWS for Windows». Это название составлено из первых букв фразы: job access with speech (доступ к рабочему месту с помощью речи). Требуемая информация озвучивается с помощью установленного синтезатора речи. Бывают как аппаратные, так и программные синтезаторы речи. Программные синтезаторы вытиснили аппаратные из-за удобства эксплуатации. Среди русских голосов большую популярность получил синтезатор «speaking mouse» («говорящая мышь»).

Главным преимуществом этого голоса является высокая скорость речи 12,3 в формате MP3 и XML-файл, созданных с помощью первого модуля.

Работа пользователя состоит из двух шагов: необходимо выбрать закладку с подходящим названием и запустить воспроизведение звука. Программа, пользуясь информацией из XML-файла, автоматически перейдет к нужному файлу и произведет смещение по времени внутри него. Книга будет звучать непрерывно до тех пор, пока пользователь не активирует другую закладку или не остановит воспроизведение.

«Говорящая книга» давно и успешно используется незрячими читателями-слушателями. При чтении (прослушивании) учебника, часто возникает необходимость оперативно ознакомиться с содержанием одного из предыдущих разделов. Иногда, например, при повторении материала, приходится конспективно прослушивать короткие фрагменты из нескольких разделов. Вообще, найдя нужный раздел по содержанию (содержание книги обычно записывается в начале или конце всей аудиокниги), надо перейти к началу самого раздела и начать слушать. Обычно любое из перечисленных действий занимает много времени. В настоящее время широкое распространение получил формат MP3 для компактного хранения звука. Несмотря на значительное преимущество компьютера (или аппаратного MP3-плеера), поиск фрагмента записи по-прежнему является достаточно длительным процессом. Для поиска приходится по очереди прослушивать несколько десятков MP3-файлов.

4. ДИСЦИПЛИНЫ ОБЛАСТИ ПРИКЛАДНОЙ ИНФОРМАТИКИ

На сегодняшний день незрячие и слабовидящие используют компьютер, как для учебы и работы, так и для отдыха: программируют, читают книги, слушают и сочиняют музыку. Для того чтобы компьютер мог служить средством реабилитации, необходимо оснастить его программами, обеспечивающими специальный интерфейс. Наиболее распространенные из них – синтезатор речи, программа экранного доступа для чтения экрана, программа оптического распознавания текста (со сканером).

Однако в любом случае слабовидящий сталкивается с проблемой обнаружения и опознания необходимой ему информации на экране дисплея, следовательно, успех или неудача его деятельности напрямую зависит от свойств пользовательского интерфейса.

Если интерфейс, как в операционной системе *Unix*, исходно представлен текстовыми строками, казалось бы, проблемы решаются не так сложно: текст можно озвучить специальной программой. Однако для ОС *Unix* как раз этого и нет! Под эту операционную систему в настоящее время разработано слишком мало специальных средств. При проведении занятий по основам администрирования ОС *Unix*, т.е. по работе с командной строкой, приходится сажать зрячего и незрячего студентов вместе, чтобы первый мог контролировать набор текста, поскольку второй вынужден все делать вслепую. Чтобы незрячий студент мог проанализировать результат своей деятельности, приходится специально ставить программу-клиент SecureCrt, записывающую сеанс работы в текстовый файл, а потом, озвучив в Windows этот текст, незрячий может «посмотреть», что он делал и какие результаты получил.

В современных средствах программирования и проектирования существенное значение имеет графическое представление информации, что ставит слушателей рассматриваемой категории в тяжелое положение. Нельзя сказать, что нет никаких средств для улучшения ситуации. Существует компонент «лупа», позволяющий увеличить фрагмент экрана, есть программы озвучивания текста. Однако этого недостаточно, да и сами компоненты неудовлетворительны. В частности, лупа занимает полосу экрана, что достаточно удобно для текста, но для представления графики лучше, если бы она увеличивала произвольную прямоугольную область. Изображение в ней состоит из квадратиков, символизирующих пиксели, что дробит изображение и делает его непонятным. Далее, такой популярный пакет как *Delphi* дает возможность создавать проекты с любым масштабом компонентов, но интерфейс среды разработки не масштабируется. То же самое можно сказать и о других инструментальных средствах.

Таким образом, для организации полноценного процесса совместного обучения и предоставления равных возможностей всем обучаемым независимо от того, в какой мере у них имеются проблемы по здоровью, необходимо решить целый ряд задач:

- выделение классов обучаемых с точки зрения ограничения возможностей и потребностей в специализированных средствах;
- определение специфики преподаваемых дисциплин с позиций их восприятия слабовидящими и незрячими;
- анализ и оценка опыта преподавания отдельных дисциплин с точки зрения соответствия результата обучения желаемому (вариант для сравнения – обучение аудитории без ограничения возможностей);
- анализ традиционных средств подачи учебной информации с точки зрения их применимости в обучении рассматриваемой категории;
- возможность реализации специфических потребностей аудитории наличными средствами (например, каким должен быть текст, чтобы он воспринимался на слух, как представлять схемы, как организовать элементы интерфейса компьютерного продукта для эффективной работы и др.);
- анализ потребностей аудитории в специфических средствах обучения;
- описание сформировавшихся принципов, методик и приемов обучения, показавших себя как эффективные;
- формулирования достигнутых результатов и постановки перспективных задач.

Как уже говорилось ранее, при переходе на дистанционный формат использование информационных технологий и платформ стало, как никогда востребовано и, конечно, учитывая вышесказанное, нужно было постараться выполнить те условия, которые необходимы при обучении студентов с нарушениями зрения.

В учебном процессе использовались все известные инструменты, взаимосвязь осуществлялась с помощью социальных сетей, почты, телефонии и других средств коммуникации («Mirapolis», «Google Meet», «Zoom», «Cisco Webex», «VK» и т.п.). Активное использование этих ресурсов позволило оценить каждый из них в образовательном процессе и выявить положительные и отрицательные стороны, увидеть насколько можно полноценно взаимодействовать в этих программах. Нужно

отметить, что все эти платформы в достаточной мере соответствуют поставленным задачам, общение в чате, запуск файлов разных форматов и т.д., но не все из них соответствуют требованиям доступности для людей с инвалидностью и других лиц с ограничениями жизнедеятельности (гост Р52872–2019 Интернет-ресурсы и другая информация, представленная в электронно-цифровой форме. Приложения для стационарных и мобильных устройств, иные пользовательские интерфейсы), а, учитывая, что на факультете обучается много слабовидящих и незрячих студентов, то такая проблема стала в период пандемии особенно актуальной.

Дистанционное обучение, в том числе для лиц с ОВЗ сопряжено и с рядом проблем. Одной из самых главных и наиболее явных проблем является коммуникация преподавателей и студентов. Условия дистанционного образования лишают студентов и преподавателей возможности очного общения, это может вызывать трудности для инклюзивного образования студентов с ОВЗ. Возникают новые коммуникативные барьеры в общении. Усвоение информации в подобной форме обучения отличается от особенностей усвоения при очном обучении. Студенты могут испытывать проблемы с усвоением материала, особенно с восприятием материала, который включает в себя формулы, графики, схемы, алгоритмы и т.п. И не стоит забывать о наличии наиболее весомой проблемы – технической подготовленности всех участников образовательного процесса.

Всемирный переход на дистанционное обучение стимулировал разработку и апробацию различных онлайн платформ («BigBlueButton», «Microsoft Teams», «Skype», «Zoom», «Cisco Webex» и др.), массовое их использование позволило выделить «слабые стороны» и совершенствовать их сообразно потребностям образовательного процесса. Поэтому большинство из них соответствуют решению всех поставленных в процессе обучения задач: наличие чата, возможности трансляции презентаций, аудио- и видеоматериалов, и пр. На факультете были использованы следующие платформы: Mirapolis LMS, Google Meet, Webex, Zoom.

Mirapolis LMS – современная система управления дистанционным образованием.

Платформа «Виртуальная комната» предназначена для организации дистанционного обучения. Может использоваться также для:

- конференций;
- совещаний;

Функционал и особенности сервиса:

- простой интерфейс, настройка шаблонов «под свой вкус»;
- поддержание разных форматов передачи информации: видео, аудио, общение в чате и в личных сообщениях;
- настраиваемый доступ, распределение ролей;
- подготовка события и простая процедура регистрации участников;
- возможность группового просмотра видеоконтента и других материалов;
- работа с документацией, в том числе совместная;
- демонстрация экрана;
- проведение опросов, голосований, тестирований;

- запись хода мероприятия и выступлений, конвертирование информации в видеоролики;
- создание отчетов, импортирование данных;
- уведомления, рассылка напоминаний по электронной почте;
- ведение статистики;
- наличие интерактивной доски, электронной указки, возможность передачи доступа к клавиатуре и мыши.

По опыту использования на факультете информационных технологий онлайн-платформа «Mirapolis LMS» полностью недоступна для незрячих пользователей. Элементы интерфейса не управляются с клавиатуры, а также не озвучиваются при сенсорном взаимодействии через мобильные устройства. Собственно, вот и всё неудобство: оно неуправляемо для пользователей программ экранного доступа.

«Google meet» входит в состав пакета услуг «Google Workspace», позволяет проводить видеоконференции, используя параллельно остальные продукты компании, которые предназначены для: хранения, обработки и создания различных данных и документов необходимых непосредственно в учебном процессе.

WebEx – платформа для проведения видеоконференций. На ней реализована поддержка аудио с помощью аналоговых телефонов и VoIP телефонии. Программу можно запустить на компьютере или смартфоне. Функционал сервиса WebEx:

- проведение веб-конференций и прямых трансляций;
- прием видео в высоком качестве;
- голосовое управление переключением камер;
- режим многопользовательского просмотра файлов;
- записи трансляций и сохранение на облачном диске;
- передача прав организатора другому участнику видео-встречи;
- командные и приватные чаты.

Сервис позволяет проводить конференции при подключении к WiFi, а также через мобильный интернет 3G/LTE.

«Zoom» предлагает своим пользователям возможность бесплатного проведения видеоконференций в рамках временного ограничения, голосовую связь, вебинары и чаты на персональном компьютере, на мобильных устройствах и в системах конференц-залов.

Преимущества Zoom:

- Видео и аудио связь с каждым участником. У организатора есть возможность выключать и включать микрофон, а также выключать видео и запрашивать включение видео у всех участников. Можно войти в конференцию как участник с правами только для просмотра.
- Можно делиться экраном (screensharing) уже со звуком. Демонстрацию экрана можно поставить на паузу. Более того, можно делится не всем экраном, а только отдельными приложениями, например, включить демонстрацию браузера. В настройках можно дать всем участникам возможность делиться экранам, либо включить ограничения, чтобы делать это мог только организатор.

- В платформу встроена интерактивная доска, можно легко и быстро переключаться с демонстрации экрана на доску.
- Есть чат, в котором можно писать сообщения, передавать файлы всем или выбрать одного студента. Чат можно настроить на автоматическое сохранение или сохранять вручную при каждой конференции (Чат → Подробнее → Сохранить чат).
- Можно производить запись урока как на компьютер, так и на облако. Удобно, что можно настроить автовключение записи, а также ставить ее на паузу.
- Во время конференции можно назначить со-организатора, у которого будут такие же возможности как и у организатора: включать и выключать микрофон у отдельных студентов, переименовывать и делить на комнаты.
- Можно как на офлайн-занятии разделить студентов и дать отдельные задания. Можно студентов разделить на пары и группы и распределить их в отдельные комнаты – сессионные залы (мини-конференции), где они будут общаться только друг с другом. Остальные их не будут ни видеть, ни слышать. Количество комнат определяет преподаватель, участников можно распределить автоматически или в ручную. У организатора есть возможность ходить по комнатам и проверять, что там происходит. Также можно перемещать участников из комнаты в комнату.

Лучшими ресурсами для обучения незрячих студентов можно назвать следующие платформы:

- Zoom (все элементы озвучиваются, озвучиваются и уведомления от программы (включение звука, начало записи, выход участника и т.д.),
- Google meet,
- Cisco Webex (неудобство поиска некоторых кнопок является делом привычки).

Все обобщающие выводы сделаны на основе опроса незрячих и слабовидящих студентов факультета ИТ.

Использования цифровых и сетевых технологий и платформ стало наиболее актуально в период пандемии COVID19, когда формат обучения приобрел исключительно дистанционный формат. Условия ограничения контактной работы и досуга вызвали ряд противоречий и трудностей, однако, параллельно, определили позитивный вектор в развитии навыков работы с сетевыми и цифровыми ресурсами и платформами как среди преподавательского состава, так и среди студентов, что вывело процесс обучения на качественно новый уровень. При этом дистанционный формат обучения довольно удобен для маломобильной категории населения, однако для обучающихся с нарушениями зрения в силу особенностей их восприятия, следует отметить ряд сложностей при участии в онлайн-лекциях, вебинарах и пр. Восприятие материала данной категории обучающихся отлично от традиционного, что подразумевает трансформацию образовательного процесса в режиме онлайн таким образом, чтобы поступающая информация была доступна, соотносима с особенностями, образовательными потребностями каждого студента. Также выполнение заданий в онлайн формате является не всегда доступным и удобным. В связи с чем, необходимо отметить ряд организационно-технологических условий, способных учесть психофизические особенности и возможности, обучающихся с ОВЗ и удовлетворить их особые образовательные потребности в условиях дистанционного обучения:

– наличие оснащенного необходимым компьютерно-программным комплексом с сетевым обеспечением, ассистивным оборудованием, методическое обеспечение в адаптированных под нозологические особенности вариациях;
– представление (дублирование) информации (лекции, семинара, практического задания и пр.) в доступном формате;
– предоставление заданий заблаговременно, учитывая временные особенности восприятия и возможности выполнения их студентами с ОВЗ;
– учет особенностей речеведения педагогами при проведении онлайн-занятия;
– обязательным является сопровождение онлайн-занятия посредством чата. Ассистивные технологии и программные средства стали особо актуальны в период дистанционного обучения. Помимо прочего, необходимо учитывать наличие компьютерной техники (включая устройства ввода и вывода), колонок, зоны для письма и дополнительного ассистивного оборудования; наличие хорошего освещения (рекомендуется теплый свет, преимущественно – естественный, либо наличие лампы); оборудование компьютерной техникой (ПК/ноутбук с вебкамерой, микрофоном, динамиками (колонками); клавиатура, мышь, устройства вывода информации и пр.). Обязательно наличие стабильного доступа к сети «Интернет»; наличие ассистивных технологий – оборудования и программно-сетевых ресурсов, способных удовлетворить особые образовательные потребности обучающихся различных нозологий. Выбор и вариативность комбинирования ассистивных технологий прямо пропорциональны особенностям и возможностям обучающегося с ОВЗ по зрению.

5. ЗАКЛЮЧЕНИЕ

«Новая коронавирусная реальность» обнажила проблемы, которые накапливались в образовании на протяжении многих лет: неразвитость инфраструктуры, недостаточная готовность многих к работе в дистанционном режиме, недостаток качественных и мощных онлайн-ресурсов и др.

Сложность выявилась и в неравенстве возможностей и условий для дистанционного обучения: наиболее мотивированные студенты легко перешли в онлайн, и этот режим оказался для них зачастую весьма комфортным, а немотивированные ребята в ряде случаев просто прекратили учиться.

Пандемия наглядно показала, что цифровизация из вспомогательного направления развития стала основным, а «виртуальная» реальность уже давно превратилась в реальное средство образования. Новые методы цифрового образования диктуют необходимость разработки новой дидактики, а это, в свою очередь, потребует перестройки не только педагогического образования, но и образа мышления всех работников образовательной системы. Опыт массового дистанционного обучения на факультете «Информационные технологии» актуализировал необходимость формирования новых методов и способов обучения в условиях цифрового общества. Это потребует организации новых форм взаимодействия участников образовательных

отношений, создания принципиально другой технологической инфраструктуры и разработки новых комплексов обеспечения информационной безопасности.

Методика и приемы хорошего дистанционного обучения направлены на то, чтобы любой студент не чувствовал себя брошенным, одиноким, изолированным от остальных. Здесь должны быть созданы все условия для продуктивной атмосферы взаимодействия. В заключении хочется отметить, что электронное образование является практически идеальным для организации дистанционного обучения лиц с ограниченными возможностями здоровья. Вопрос заключается только в том, чтобы электронное образование не вытеснило традиционное образование, а интегрировалось в него.

Литература

1. Федеральный закон «Об образовании в РФ» от 29 декабря 2012 г. № 273-ФЗ.
2. Федеральный закон «О внесении изменений в статью 71 Федерального закона «Об образовании в Российской Федерации» от 1 мая 2017 года № 93-ФЗ.
3. Федеральный закон «О внесении изменений в отдельные законодательные акты Российской Федерации по вопросам социальной защиты инвалидов в связи с ратификацией Конвенции о правах инвалидов». от 1 декабря 2014 года № 419-ФЗ.
4. Федеральный закон № 149-ФЗ «Об информации, информационных технологиях и о защите информации» от 27.07.2006.
5. Федеральный закон № 152-ФЗ «О персональных данных» от 27.07.2006.
6. Приказ Минобрнауки РФ от 23 августа 2017 г. № 816 «Об утверждении Порядка применения организациями, осуществляющими образовательную деятельность, электронного обучения, дистанционных образовательных технологий при реализации образовательных программ».
7. Приказ Минобрнауки РФ от 5 апреля 2017 г. № 301 «Об утверждении Порядка организации и осуществления образовательной деятельности по образовательным программам высшего образования – программам бакалавриата, программам специалитета, программам магистратуры».
8. Постановление Правительства Российской Федерации от 17 марта 2011 г. N 175 г. Москва «О государственной программе Российской Федерации «Доступная среда» на 2011–2015 годы».
9. Приказа Минобрнауки от 16.04.2014 г. № 05–785 «О направлении методических рекомендаций по организации образовательного процесса для обучения инвалидов».
10. Методических рекомендаций по организации образовательного процесса для обучения инвалидов и лиц с ограниченными возможностями здоровья в образовательных организациях высшего образования, в том числе оснащенности образовательного процесса» (утв. Минобрнауки России 08.04.2014 N АК-44/05вн).
11. *Нуркаева И.М.* Особенности обучения программированию незрячих студентов МГППУ образовании. Сб. науч. трудов. – М.: МИФИ, 2004. – Ч. IV. – С. 100–101.
12. *Нуркаева И.М., Коморина К.А.* Информационная система диагностики профессионального выгорания педагогов // Моделирование и анализ данных. – М.: ФГБОУ ВО МГППУ, 2017. – Т. 1 – № 1 – С. 95–103.
13. *Нуркаева И.М., Зайцев А.Н., Оглоблин А.А.* Информационная система для мониторинга учебных достижений студентов МГППУ // Моделирование и анализ данных. – М.: ФГБОУ ВО МГППУ, 2019 – № 1 – С. 30–41.

14. *Шефер Е.А.* Использование цифровых технологий в образовательном процессе / Е.А. Шефер. – Текст : непосредственный // Молодой ученый. – 2021. – № 16 (358). – С. 22–25. – URL: https://moluch.ru/archive/358/79973/ (дата обращения: 17.01.2022).
15. Электронный ресурс. URL: https://www.profguide.io/article/chem-otlichaetsya-distancionnoe-obuchenie-ot- ehlektronnogo-i-onlajn.html

Features of Teaching Students with Visual Disabilities in the Disciplines of Mathematical and Computer Cycles at the Faculty of Information Technology with the Use of Remote Technologies

Elena B. Cherven-Vodali *
Moscow state University of Psychology & Education (MSUPE), Moscow, Russia
ORCID: https://orcid.org/0000-0002-6871-9105
e-mail: cervenvodali@mgppu.ru

Svetlana N. Antipova **
Moscow state University of Psychology & Education (MSUPE), Moscow, Russia
ORCID: https://orcid.org/0000-0001-6642-7953
e-mail: antipovasn@mgppu.ru

Valeriya B. Sidorova ***
Moscow state University of Psychology & Education (MSUPE), Moscow, Russia
ORCID: https://orcid.org/0000-0001-6391-5361
e-mail: sidorovavb@mgppu.ru

The article deals with the organization of the educational process in the conditions of distance learning for students with disabilities and persons with visual disabilities in connection with the introduction of self-isolation during the period of the incidence of Covid19. Special emphasis is placed on teaching computer and mathematical cycle disciplines to students with visual impairment, as these disciplines have their own specifics.

Keywords: educational process, distance learning technologies, students with disabilities.

For citation:
Cherven-Vodali E.B., Antipova S.N., Sidorova V.B. Features of Teaching Students with Visual Disabilities in the Disciplines of Mathematical and Computer Cycles at the Faculty of Information Technology with the Use of Remote Technologies. *Modelirovanie i analiz dannykh = Modelling and Data Analysis,* 2022. Vol. 12, no. 1, pp. 60–78. DOI: https://doi.org/10.17759/mda.2022120105 (In Russ.,abstr. in Engl.).

***Elena B. Cherven-Vodali,** Lecturer of the Department of Applied Informatics and Multimedia Technologies, Moscow state University of Psychology & Education (MSUPE), Moscow, Russia, ORCID: https://orcid.org/0000-0002-6871-9105, e-mail: cervenvodali@mgppu.ru

****Svetlana N. Antipova,** Deputy Dean for Extracurricular Activities of the Faculty of Information Technology, Moscow state University of Psychology & Education (MSUPE), Moscow, Russia, ORCID: https://orcid.org/0000-0001-6642-7953, e-mail: antipovasn@mgppu.ru

*****Valeriya B. Sidorova,** Lecturer of the Department of Applied Informatics and Multimedia Technologies, Moscow state University of Psychology & Education (MSUPE), Moscow, Russia, ORCID: https://orcid.org/0000-0001-6391-5361, e-mail: sidorovavb@mgppu.ru

References

1. Federal'nyi zakon «Ob obrazovanii v RF» ot 29 dekabrya 2012 g. № 273-FZ.
2. Federal'nyi zakon «O vnesenii izmenenii v stat'yu 71 Federal'nogo zakona «Ob obrazovanii v Rossiiskoi FederatsiI» ot 1 maya 2017 goda № 93-FZ.
3. Federal'nyi zakon «O vnesenii izmenenii v otdel'nye zakonodatel'nye akty Rossiiskoi Federatsii po voprosam sotsial'noi zashchity invalidov v svyazi s ratifikatsiei Konventsii o pravakh invalidoV». ot 1 dekabrya 2014 goda № 419-FZ.
4. Federal'nyi zakon № 149-FZ «Ob informatsii, informatsionnykh tekhnologiyakh i o zashchite informatsiI» ot 27.07.2006.
5. Federal'nyi zakon № 152-FZ «O personal'nykh dannyKH» ot 27.07.2006.
6. Prikaz Minobrnauki RF ot 23 avgusta 2017 g. № 816 «Ob utverzhdenii Poryadka primeneniya organizatsiyami, osushchestvlyayushchimi obrazovatel'nuyu deyatel'nost', ehlektronnogo obucheniya, distantsionnykh obrazovatel'nykh tekhnologii pri realizatsii obrazovatel'nykh programM».
7. Prikaz Minobrnauki RF ot 5 aprelya 2017 g. № 301 «Ob utverzhdenii Poryadka organizatsii i osushchestvleniya obrazovatel'noi deyatel'nosti po obrazovatel'nym programmam vysshego obrazovaniya – programmam bakalavriata, programmam spetsialiteta, programmam magistraturY».
8. Postanovlenie Pravitel'stva Rossiiskoi Federatsii ot 17 marta 2011 g. N 175 g. Moskva «O gosudarstvennoi programme Rossiiskoi Federatsii "Dostupnaya sreda" na 2011–2015 gody»
9. Prikaza Minobrnauki ot 16.04.2014 g. № 05–785 «O napravlenii metodicheskikh rekomendatsii po organizatsii obrazovatel'nogo protsessa dlya obucheniya invalidoV»/
10. Metodicheskikh rekomendatsii po organizatsii obrazovatel'nogo protsessa dlya obucheniya invalidov i lits s ogranichennymi vozmozhnostyami zdorov'ya v obrazovatel'nykh organizatsiyakh vysshego obrazovaniya, v tom chisle osnashchennosti obrazovatel'nogo protsessa" (utv. Minobrnauki Rossii 08.04.2014 N AK-44/05vn).
11. Nurkaeva I.M. Osobennosti obucheniya programmirovaniyu nezryachikh studentov MGPPU obrazovanii. Sb. nauch. trudov. – M.: MIFI, 2004 – ch. IV. – S. 100–101.
12. Nurkaeva I.M., Komorina K.A. Informatsionnaya sistema diagnostiki professional'nogo vygoraniya pedagogov . Modelirovanie i analiz dannykh. – M.: FGBOU VO MGPPU, 2017 – T. 1 – № 1 – S. 95–103.
13. Nurkaeva I.M., Zaitsev A.N., Ogloblin A.A. Informatsionnaya sistema dlya monitoringa uchebnykh dostizhenii studentov MGPPU . Modelirovanie i analiz dannykh. – M.: FGBOU VO MGPPU, 2019 – № 1 – S. 30–41.
14. Shefer, E.A. Ispol'zovanie tsifrovykh tekhnologii v obrazovatel'nom protsesse . E.A. Shefer. – Tekst : neposredstvennyi . Molodoi uchenyi. – 2021. – № 16 (358). – S. 22–25. – URL: https://moluch.ru/archive/358/79973/ (data obrashcheniya: 17.01.2022).
15. Ehlektronnyi resurs.URL: https://www.profguide.io/article/chem-otlichaetsya-distancionnoe-obuchenie-ot- ehlektronnogo-i-onlajn.html (data obrashcheniya: 28.03.2022).

Получена 02.02.2022
Принята в печать 16.02.2022

Received 02.02.2022
Accepted 16.02.2022

Моделирование и анализ данных 2022. Том 12. № 1.
Научный журнал

Издаётся с 2011 года

Учредитель
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Московский государственный психолого-педагогический университет»

Адрес редколлегии:
г. Москва, ул. Сретенка, 29, факультет информационных технологий
Тел.: +7 (499) 167-66-74
E-mail: mad.mgppu@gmail.com

Журнал зарегистрирован в Государственном комитете РФ по печати.
Свидетельство о регистрации средств массовой информации
ПИ № ФС77-52058 от 7 декабря 2012 года

ISSN: 2219-3758
ISSN: 2311-9454 (online)

Подписано в печать: 11.03.2022.
Формат: 70*90/16. Бумага офсетная.
Гарнитура Times. Печать цифровая.
Усл. печ. п. 4,2. Усл.-изд. л. 4,9.
Тираж 500 экз.