

Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.85:81'243

Сафоненко
Карина Андреевна

«Алгоритмы повышения эффективности обучения
иностранному языку на основе методов машинного обучения»

Автореферат

на соискание степени магистра технических наук
по специальности 1-40 80 02 «Системный анализ, управление
и обработка информации»

Научный руководитель
Бойко И. М.
кандидат технических наук

Минск 2021

КРАТКОЕ ВВЕДЕНИЕ

Существуют различные причины изучения иностранных языков: карьерный рост, получение образования за рубежом и получение вида на жительство, просмотр фильмов и чтение книг на языке оригинала и многие другие. Однако в современном ритме жизни людям не всегда удается найти достаточное количество времени для обучения и даже те, кто его находят, не всегда приходят к желаемому результату. Таким образом, повышение эффективности обучения иностранным языкам становится все более актуальной задачей в современном мире. Для решения этой задачи необходимо придерживаться методик, которые могли бы соответствовать физиологическим особенностям человека, а также отвечать его ритму жизни.

Образование является сферой, в которой производится и накапливается большое количество данных:

- Образовательный процесс подразумевает под собой сотни или тысячи часов, проведенных в учебном заведении, на онлайн-платформе или за экраном телефона в течение продолжительного промежутка времени. Такое взаимодействие обучаемого с учебными материалами генерирует множество информации.

- В системах управления образованием (LMS – Learning Management Systems), образовательных онлайн-платформах или приложениях по изучению иностранных языков собираются данные о взаимодействии студента с системой, о его прогрессе и результатах выполнения заданий и упражнений.

Правильный анализ такой информации может помочь получить более полную картину процесса обучения отдельного учащегося и выявить, возможно, неочевидные связи, которые в дальнейшем могут быть использованы для повышения эффективности обучения.

Машинное обучение представляет собой раздел искусственного интеллекта, использующий методы построения алгоритмов, способных обучаться в процессе применения решений множества сходных задач. Область применения методов машинного обучения постоянно расширяется и затронула практически все сферы человеческой деятельности, в том числе и сферу образования. В отличие от традиционных методов анализа результатов учащихся, которые измеряют только конечный результат учащегося, применение методов машинного обучения может помочь получить ценную информацию о том, как улучшить и персонализировать обучение, основываясь на данных взаимодействия конкретного пользователя с системой.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Цель и задачи исследователя

Цель данной работы – демонстрация возможностей применения машинного обучения в сфере образования, а также разработка алгоритмов повышения эффективности обучения иностранным языкам на основе методов машинного обучения.

Для достижения поставленной цели научного исследования были решены следующие задачи:

- проанализированы существующие методы машинного обучения и сферы их применения;
- проанализированы существующие компьютерные технологии обучения иностранным языкам;
- сформулированы ограничения существующих систем обучения иностранным языкам;
- проанализированы способы применения методов машинного обучения в сфере образования при различных формах образования: традиционной (аудиторной), смешанной (аудиторной и удалённой) и дистанционной;
- сформулирована задача на разработку рекомендательных алгоритмов по подбору иностранных слов для изучения с учетом индивидуальных особенностей и потребностей каждого учащегося;
- произведена генерация синтетического набора данных для тренировки и тестирования разрабатываемых алгоритмов машинного обучения;
- разработаны рекомендательные алгоритмы по подбору иностранных слов для изучения на основе метода «k-средних» (k-means) и определения количества кластеров по «методу локтя» (WCSS);
- выполнена реализация предлагаемых алгоритмов с использованием языка программирования Python, в составе специализированного дистрибутива для научных вычислений Anaconda, среды разработки Jupyter Notebooks и библиотек для работы с данными и моделями машинного обучения NumPy и Scikit-learn;
- приведена оценка качества работы предложенных алгоритмов по метрикам: «аккуратности» (accuracy), точности (precision), полноты (recall) и F1.

Объектами научного исследования являются компьютерные технологии обучения иностранным языкам, методы машинного обучения.

Предметом научного исследования является повышение эффектив-

ности обучения иностранным языкам на основе методов машинного обучения.

Новизна полученных результатов

Новизна полученных результатов диссертационной работы заключается в определении подходов повышения эффективности обучения иностранным языкам с использованием методов машинного обучения, разработке и реализации рекомендательных алгоритмов по подбору иностранных слов для их изучения с учетом индивидуальных особенностей учащегося.

Положения, выносимые на защиту

- рекомендательный алгоритм по подбору иностранных слов для изучения, реализованный на основе кластерного анализа, учитывающий пользовательский уровень владения иностранным языком;
- рекомендательный алгоритм по подбору иностранных слов для изучения, реализованный на основе кластерного анализа, учитывающий пользовательский уровень владения иностранным языком и интересы пользователя.

Апробация результатов

Результаты диссертационной работы обсуждались на конференции «Информационные технологии и системы 2020» (БГУИР, Минск, Беларусь, 2020).

Опубликованность результатов исследования

По теме диссертации опубликована 1 печатная работа в сборнике материалов международной конференции.

Структура и объем диссертации

Диссертация состоит из введения, четырех глав, заключения, списка публикаций автора и списка использованных источников. Общий объем магистерской диссертации составляет 66 страниц, включая 36 иллюстраций, 1 таблицу и список использованных источников из 56 наименований.

КРАТКОЕ СОДЕРЖАНИЕ

Во **введении** обоснована актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначены актуальные задачи, решению которых посвящена диссертационная работа.

В **первой главе** были рассмотрены подходы к определению термина «Машинное обучение».

Также были рассмотрены следующие виды машинного обучения:

- Обучение с учителем;
- Обучение без учителя;
- Обучение с частичным привлечением учителя;
- Обучение с подкреплением.

Среди задач, решаемых с помощью методов машинного обучения были выделены:

- Регрессия;
- Классификация;
- Кластеризация;
- Снижение размерности;
- Выявление аномалий.

Были также описаны используемые алгоритмы для решения вышеперечисленных задач и сферы их применения в реальной жизни.

Во **второй главе** внимание было уделено компьютерным технологиям обучения иностранным языкам.

Были описаны основные преимущества применения компьютерных технологий в обучении, среди которых были выделены следующие:

- оперативная передача и представление информации любого объема и вида;
- хранение этой информации в памяти компьютера в течение необходимой продолжительности времени, возможность ее редактирования, обработки и т.д.;
- интерактивность с помощью специальных аппаратных средств;
- доступ к различным источникам информации, удаленным базам данных, работа с этой информацией;
- возможность организации электронных аудио и видеоконференций, игр, в том числе в режиме реального времени и многие другие факторы.

Был произведен обзор существующих методов (с акцентом на лексику, фонетику, грамматику языка) и программных средств обучения иностранным языкам, таких как «Duolingo», «Lingualeo», «Memrise».

Были также сформулированы ограничения существующих компьютерных систем обучения, среди которых были выделены:

- недостаточная интерактивность (содержательную основу многих программ составляют текстовые материалы и простейшие графические объекты (рисунки, фото), блоки контроля знаний оформлены в виде тестовых текстовых заданий);
- отсутствие единой универсальной методики использования информационно-коммуникационных технологий в обучении иностранным языкам;
- ограниченные возможности баз данных для хранения учебного материала;
- отсутствие индивидуального подхода к обучающемуся;
- снижение концентрации учебного внимания (вместо осознания и осмысления нового материала, оказание большего внимания на эффекты информационно-коммуникационных технологий).

В **третьей главе** был описан процесс формирования и накопления данных в традиционных учебных заведениях, таких как школы, колледжи, высшие учебные заведения, и в системах онлайн образования.

Был произведен обзор применения методов машинного обучения в сфере образования при различных формах образования: традиционной (аудиторной), смешанной (аудиторной и удалённой) и дистанционной. Были также рассмотрены реальные примеры и результаты применения данных методов в различных учебных заведениях.

В **четвертой главе** была сформулирована задача разработки рекомендательных алгоритмов по подбору иностранных слов для изучения с учетом индивидуальных особенностей и потребностей каждого учащегося.

Для решения данной задачи была произведена генерация синтетического набора данных для тренировки и тестирования разрабатываемых алгоритмов машинного обучения, по основным признакам, используемым в системе.

В результате были разработаны алгоритмы кластеризации на основе метода «к-средних» (k-means) и определения количества кластеров по «методу локтя» (WCSS). Реализация предлагаемых алгоритмов была выполнена с использованием языка программирования Python, в составе спе-

специализированного дистрибутива для научных вычислений Anaconda, среды разработки Jupyter Notebooks и библиотек для работы с данными и моделями машинного обучения NumPy и Scikit-learn.

Для первого алгоритма были выделены следующие параметры:

- возраст пользователя;
- количество добавленных пользователем слов из категории «Общая формальная лексика»;
- количество добавленных пользователем слов из категории «Профессиональная лексика»;
- количество добавленных пользователем слов из категории «Официально-деловая лексика»;
- количество добавленных пользователем слов из категории «Газетно-публицистическая лексика».

В результате работы алгоритма были получены три кластера пользователей.

Пользователи, принадлежащие к первому кластеру:

- Добавляли наибольшее количество слов из категории «Общая формальная лексика»;
- Добавляли наименьшее количество слов из категории «Профессиональная лексика»;
- Добавляли наименьшее количество слов из категории «Официально-деловая лексика»;
- Добавляли мало слов из категории «Газетно-публицистическая лексика».

Пользователи, принадлежащие ко второму кластеру:

- Добавляли среднее количество слов из категории «Общая формальная лексика»;
- Добавляли наибольшее количество слов из категории «Профессиональная лексика»;
- Добавляли среднее количество слов из категории «Официально-деловая лексика»;
- Добавляли мало слов из категории «Газетно-публицистическая лексика».

Пользователи, принадлежащие к третьему кластеру:

- Добавляли наименьшее количество слов из категории «Общая формальная лексика»;

- Добавляли среднее количество слов из категории «Профессиональная лексика»;
- Добавляли наибольшее количество слов из категории «Официально-деловая лексика»;
- Добавляли наибольшее количество слов из категории «Газетно-публицистическая лексика».

Если проводить аналогию между полученными кластерами и уровнями владения английского языка (для примера), то можно сделать следующие выводы:

- первый кластер соответствует уровням A1, A2 (где акцент в изучении стоит на понимании и использовании повседневных выражений, основных фраз, направленных на удовлетворение базовых потребностей);
- второй кластер - уровням B1, B2 (где акцент в изучении стоит на развитие умения поддерживать диалог на более сложные темы);
- третий кластер - уровням C1, C2 (где акцент в изучении ставится на свободное выражение своего мнения на любую тему, даже на узко ориентированную).

Ко второму алгоритму были добавлены дополнительные параметры:

- количество добавленных пользователем слов из категории «Компьютерные игры»;
- количество добавленных пользователем слов из категории «Спорт»;
- количество добавленных пользователем слов из категории «Программирование»;
- количество добавленных пользователем слов из категории «Экономика».

В результате работы данного алгоритма были получены пять кластеров пользователей. В отличие от кластеров, полученных в результате работы первого алгоритма, данные кластеры учитывают не только возраст пользователя и распределение пользователей по добавлению слов из различных категорий лексики (или в данном случае распределению пользователей по уровням владения языком), но и интересы каждого отдельного пользователя.

В этой же главе была приведена оценка качества работы предложенных алгоритмов по метрикам: ассурасу, точности (precision), полноты (recall) и F1.

ЗАКЛЮЧЕНИЕ

Машинное обучение представляет собой раздел искусственного интеллекта, использующий методы построения алгоритмов, способных обучаться в процессе применения решений множества сходных задач. Область применения методов машинного обучения постоянно расширяется и затронула практически все сферы человеческой деятельности, в том числе и сферу образования.

Современные компьютерные технологии обучения иностранным языкам не лишены недостатков, среди которых можно выделить отсутствие индивидуального подхода к каждому пользователю. Анализ информации с использованием методов машинного обучения, полученной в процессе использования обучающей системы, позволит повысить эффективность процесса обучения, что является актуальной задачей в связи с повышением необходимости освоения иностранного языка в кратчайшие сроки.

В данной работе основное внимание было уделено повышению эффективности пополнения словарного запаса учащегося, являющегося основой при изучении любого иностранного языка. Для решения этой задачи были разработаны рекомендательные алгоритмы по подбору иностранных слов для изучения с учетом индивидуальных особенностей и потребностей каждого учащегося. Данные алгоритмы с помощью кластерного анализа формирует список рекомендаций слов для изучения каждому отдельному пользователю. Данный список слов формируется с учетом интересов пользователя.

Дальнейшее развитие данных алгоритмов предполагает тестирование алгоритмов на реальных пользователях, тестирование других методов кластеризации и расширение набора исходных данных для получения более точных рекомендаций.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Сафоненко, К. А. Адаптивная система обучения английскому языку на основе кластерного анализа / К. А. Сафоненко, П. Д. Войтиховский, И. Т. Давыденко // Информационные технологии и системы 2020 (ИТС 2020) = Information Technologies and Systems 2020 (ITS 2020) : материалы междунар. науч. конф. (Республика Беларусь, Минск, 18 ноября 2020 года) / редкол. : Л. Ю. Шилин [и др.]. – Минск : БГУИР, 2020. – С. 68 - 69.

Библиотека БГУИР