УДК 004.891.3

# РАСПОЗНАВАНИЕ АВТОРСКОГО СТИЛЯ ПУБЛИКАЦИИ С ПОМОЩЬЮ НЕЙРОННОЙ СЕТИ

**Терехов В.И., Лободенко Е.И.**

*ФГБОУ ВО «Тюменский индустриальный университет», Тюмень, e-mail: lobodenkoei@tyuiu.ru*

Проблема заимствований текстов в последнее время стоит очень остро, поэтому для подтверждения авторства работ обучающихся и сотрудников вуза была создана нейронная сеть, распознающая стиль написания статей автора. Под понятием «авторский стиль» понимается отнесение текста к определенной науке рассматриваемой статьи, использование автором присущих только ему определенных структур построения предложений, фразеологических оборотов речи, неологизмов, вводных слов и определенных словосочетаний. В работе представлен алгоритм построения такой сети, включающий четыре этапа. На первом этапе происходит чтение и классификация текста по принадлежности к определенной науке. На втором – анализ работ автора статьи и определение процента самоцитируемости. Далее проводится сравнение со статьями других авторов внутри базы знаний вуза и выстраиваются возможные связи последующей коллаборации. На завершающем этапе статья, прошедшая рассмотрение и подтвержденная как уникальная, дополняет университетскую базу знаний и автоматически используется для дальнейшего обучения нейросети. В работе представлены подходы, на основе которых делается анализ качества публикаций сотрудников вуза и оказывается помощь в повышении их публикационной активности.

**Ключевые слова: анализ текста, информационные технологии, цифровизация, нейронная сеть, библиотека публикаций, публикационная активность**

# RECOGNITION OF AUTHOR'S PUBLISHING STYLE USING A NEURAL NETWORK

**Terekhov V.I., Lobodenko E.I.**

*Industrial University of Tyumen, Tyumen, e-mail: lobodenkoei@tyuiu.ru*

The problem of borrowing texts has recently become very acute, therefore, to confirm the authorship of the works of students and university staff, a neural network was created that recognizes the author's writing style. The concept of «author's style» refers to the attribution of the text to a certain science of the article in question, the use by the author of certain structures of sentence construction, phraseological turns of speech, neologisms, introductory words and certain phrases inherent only to him. The paper presents an algorithm for constructing such a network, which includes four stages. At the first stage, the text is read and classified according to its belonging to a particular science. On the second – analysis of the author's works and determination of the percentage of self-citation. Further, a comparison is made with articles by other authors within the knowledge base of the university, and possible links for subsequent collaboration are built. At the final stage, the article, which has been reviewed and confirmed as unique, supplements the university knowledge base, and is automatically used for further training of the neural network. The paper presents approaches on the basis of which an analysis of the quality of publications of university staff is made and assistance is provided to increase their publication activity.

**Keywords: text analysis, information technologies, digitalization, neural network, publication library, publication activity**

Первая нейронная сеть на базе пока единственного нейрокомпьютера Mark-1 Френка Розенблатта имеет более чем полувековую историю, хотя незаслуженно была предана забвению на несколько десятилетий [1]. Сегодня очень модным стало направление в науке по созданию и развитию искусственных нейронных сетей на основе аналоговых компьютеров [2–4], создана целая их теория [5–7]. Широко используются различные инновации, которые позволяют работать с разными базами данных, электронных библиотек, энциклопедий [8, 9]. Вместе с этим возникла необходимость быстро извлекать из них информацию простым пользователям. А это привело к «CopyandPaste» стилю (копируй и вставляй) написания как учебных, так и научных работ во всем мире.

Ученые забили тревогу, так как лавиной стали перепечатываться статьи. Академическое мошенничество стало нормой в учебном процессе, а заимствования – в научных трудах и диссертациях. Добросовестные ученые стали создавать и продвигать сервисы по выявлению неправомочных перепечаток и откровенного плагиата интеллектуального труда. Пришлось разрабатывать технологии подтверждения авторства. Программно-аппаратный комплекс «Антиплагиат, творите собственным умом» стал первым и пока лучшим в странах СНГ для проверки текстовых документов на наличие заимствований. Хотя его создатели до сих пор говорят, что проверяется только текст, таблицы и рисунки не сравниваются. Обязательно включение в процесс анализа текста человека-эксперта, который должен принимать решение о заимствованиях и нести ответственность за справку об оригинальности текста.

Современный мир, пронизанный интернетом, упрощает доступ к информации

и позволяет работать с огромной базой источников. Устройства, подключенные к интернету, позволяют любому пользователю окунуться в мир офлайн- и онлайн-библиотек, содержащих огромный кладезь книг, журналов и газет, обучающих лекций и методических указаний. Это выводит образование на новый уровень самостоятельности в обучении. Вместе с этим ребром встал вопрос о неправомерных заимствованиях в научном мире и академической недобросовестности в вузах, ведь стали доступными не только монографии и научные статьи, но и рефераты, курсовые и дипломные работы. До тех пор пока каждый вуз не будет воспитывать у обучающихся и сотрудников культуру написания собственных работ, преодолеть копирование учебных заданий и плагиат в научных статьях невозможно [10].

Сейчас, в отличие от прошлых лет, написание курсовых работ и рефератов не требует посещения библиотеки в физическом ее понимании, достаточно обратиться к системам поиска в интернете. А ощущение миллениумов с детства, что информация в интернете ничья, и отсутствие воспитания моральных норм поведения приводят к непониманию проблемы недобросовестных заимствований. Для них нормальным считается простое копирование информации из одного или нескольких источников, когда весь материал просматривается бегло и не всегда полностью соответствует изучаемой теме. Если в вузе не пресекать такой подход к представлению рефератов, курсовых и выпускных квалификационных работ обучающихся, то, став сотрудниками, они будут считать его этически возможным для себя.

Защита авторских прав и сохранение конфиденциальности информации сильно пострадали с развитием цифровых технологий. В последнее время все чаще возникают курьезные случаи, когда в отчете программа «Антиплагиат, твори собственным умом» выдает заимствования профессором или доцентом у студента, который выставил в интернет лекции своих преподавателей раньше, чем те решили их издать. Случается это от того, что данная программа находит только текстовое совпадение. Как говорят разработчики, принимать решение по каждому отдельному вопросу должен человек. Просматривая найденные машиной совпадения, эксперт определяет, включать их в отчет или нет.

Для публикации научными сотрудниками, преподавателями и студентами своих научных трудов предусматривается процедура проверки на антиплагиат отправляемых работ. Но не всегда низкий процент оригинальности – результат плохой работы авторов. Заимствования возможны при использовании общеупотребительных терминов и слов, в результате схожести мыслей, работы с одними источниками и так далее. В такой ситуации помощью может стать система на основе нейронной сети, которая будет обучена на предыдущих работах автора и сможет выделять его уникальный стиль, а не просто находить похожие слова и словосочетания. Под авторским стилем здесь понимается использование автором присущих ему определенных фразеологических оборотов речи, вводных слов, словосочетаний, определенной структуры построения предложений, отнесение к определенной науке. Для учебников и учебных пособий выделить стиль написания текста крайне сложно из-за необходимости пользоваться определениями и понятиями, устоявшимися веками. Но при описании примеров, их объяснении, доказательстве теорем можно уловить авторские филологические предпочтения. К сожалению, каждая нейронная сеть пока обучается под определенный вид деятельности. В этом пока недостаток современных подходов.

Предлагаемая нами разработка нацелена на создание базы знаний организации на основе опубликованных работ сотрудников (статей, монографий, учебных пособий и т.д.) и библиотеки рефератов, курсовых и выпускных квалификационных работ обучающихся. База регулярно обновляется и пополняется новой информацией, которая одновременно используется и для дальнейшего обучения нейронной сети, и в качестве базы знаний. Предлагаемые нами средства сервиса системы антиплагиат позволяют осуществлять сбор и анализ данных с возможностью извлечения и обработки информации об авторе, его научном направлении, стиле написания статей и монографий. Такая система после завершения и апробации разработки будет внедрена в Тюменском индустриальном университете для оказания помощи сотрудникам в их публикационной деятельности.

**Материалы и методы исследования**

Современная информационная среда – это сложная система, изменяющаяся под воздействием многих факторов от деятельности человека. Антропогенное влияние приводит к изменению условий, в которых протекают информационные процессы, трансформированные процессы меняют объекты этих процессов, модифицируя информационные системы.

Информационная среда включает в себя электронные ресурсы, которые в процессе

их использования генерируют связи, взаимосвязи, отношения. В ней образуются иерархические цепочки, определяются законы взаимодействия элементов и, как результат работы с этой средой, на выходе имеется хранение, сбор, доработка и передача информации. Сегодня все сферы нашей жизни развиваются под воздействием форсированного внедрения информационных технологий в большинство сфер жизнедеятельности человека.

Проведенный анализ литературных источников по проблеме обработки больших потоков и получения новых знаний из разнотипных данных [11–13] позволил выделить основные направления [14] развития информационных технологий на ближайшее будущее при условии осуществления сквозных решений:-

– применение IT-технологий в области мобильных приложений;

– развитие с их помощью социальных коммуникаций;

– использование облачных технологий для связи и взаимодействия различных областей деятельности человека;

– развертывание баз больших данных различной природы, их обработка и аналитика на основе предсказаний;

– совершенствование искусственного интеллекта на базах больших данных с применением машинного обучения;

– обеспечение кибербезопасности информационной среды;

– создание и усовершенствование интернета вещей в индустрии (IoT).

Образовательную среду вуза, работающую в дистанционном режиме, можно отнести к информационной системе с большими базами очень разнородных данных. И в каждом из указанных актуальных направлений необходимо развивать ИТ вуза, особенно в настоящей сложной эпидемиологической обстановке. Использование электронных библиотек и возможность быстро извлекать из них информацию привели к тому, что увеличилось академическое мошенничество. Работы с неправомерными заимствованиями потекли потоком. Возникла необходимость в создании технологии подтверждения авторства не только по сравнению отдельных слов, словосочетаний и текстовых кусков, а и по стилю написания текстов. Средствами сервиса антиплагиата появилась возможность сбора данных и их анализа с возможностью извлекать и обрабатывать дополнительную нужную информацию об авторе и сравнивать его работы в соответствии с направлением в науке и педагогической деятельностью, устанавливать его стиль, чтобы исключать случайные совпадения текстов.

## Результаты исследования и их обсуждение

Для активизации публицистической деятельности сотрудников и обучающихся вуза, а также повышения качества их работ создается интеллектуальная система на основе нейронной сети. Во-первых, система анализа научных статей должна уметь распознавать и классифицировать ее по принадлежности к определенной науке. Во-вторых, производить анализ предыдущих работ автора статьи, из которого выдается процент его самоцитируемости. В-третьих, проводится сравнение на похожесть его научного исследования со статьями других авторов внутри базы знаний вуза, выстраиваются связи. В заключение, прошедшая анализ работа и подтвержденная как уникальная статья, дополняет базу знаний университета (рис. 1).


*Рис. 1. Схема работы системы антиплагиат*

Определение принадлежности исследуемого текста к той или иной науке происходит на основе обучающего дата сета по различным областям науки. Для этого написан и внедрен нейронный сетевой алгоритм, позволяющий определять качество статьи и проводить подтверждение направления научных исследований автора. На данном этапе нейросеть выводит процентное соотношение принадлежности работы к виду данного научного направления, параллельно указывает автору смежные науки, по которым возможны научные взаимодействия и создание коллабораций [2].

Распределив веса на обучающем дата сете, система подключает алгоритм чтения и запускает вычислительную процедуру по выбору типа структур системы. Тип структуры является динамической основой нейронной сети, которые функционируют отрезками в процессе работы. Сеть, подобно человеческому мозгу, сохраняет все итерации своей работы и обучается на них посредством обучающего алгоритма.

Предлагаемая система написана на языке программирования JAVA и имеет следующую архитектуру.

1. Neiron – POJO класс, содержащий в себе информацию о значении нейрона и массы его связей, имеет методы для взаимодействия с ним (получением и установкой информации). Класс был написан так, чтобы поддерживать модификации для возможности проводить инновационные исследования.

2. Layer – класс, в котором содержатся адреса объектов нейронов. Позволяет удобно с алгоритмической точки зрения взаимодействовать с нейронами и получать информацию об их состоянии.

0. Neiro – класс, отвечающий за построение и обработку модели.

3.1. Класс собирает конфигурационные данные через интерфейс, легко интегрируется в любой проект с Java Virtual Machine и может быть легко внедрен в любые системы посредством сетевого взаимодействия.

3.2. Реализована система инициализации после получения команды так, что позволяет удобно манипулировать моделями и не перегружать машину до старта исполнения алгоритма.

3.3. Класс успешно создает и рассчитывает модель «Перцептрон», используемую в моделях, понимающих контекст использования слова.

3.4. Позволяет создавать и рассчитывать модели любого размера.

0. DataSort – класс, отвечающий за разбиение текста на слова для последующей работы. Возвращает считанный из .txt файла набор слов, удобный для дальнейшей обработки.

DB (Data Base) – класс, отвечающий за связь приложения нейронной сети с базой данных для записи и чтения значений, полученных при сборе дата сета.

TextModule – класс, комбинирующий работу «DB» и «DataSort». Предоставляет удобный интерфейс для работы с этими классами и расширяет функции нейронной сети.

Использование нейросетевой структуры позволяет проводить анализ научных трудов на основе семантики по ряду классификаторов. А также накапливать базу знаний вуза и предлагать варианты по созданию научных коллабораций как внутри, так и вне организации на основе результатов обработки текстов научных статей.

Анализ работ [8, 11, 14] позволил выстроить процесс создания и тестирования предлагаемой системы поэтапно: при обучении нейронной сети, выборе архитектуры, распределении весовых коэффициентов и её тестировании.

На первом этапе при разработке нейронной сети был сделан упор на математическое описание – это выбор типа структуры.

На втором этапе осуществления алгоритма обучения и калибровки в соответствии с архитектурой слоев проверяется эффективность распределения весов связей (рис. 2) [2].

На завершающем этапе цикла системы происходит тестирование за счет увеличения регистрации параметров входа и показателей с информационных систем хранения статей вуза, в реальном времени управляющих воздействий.

Разработан прототип системы классификации текста на основе нейронной сети. Предложенное программное обеспечение (ПО) обеспечивает базовую классификацию текста по принадлежности к определенной науке на базе ранее изученных нейросетью данных. В основу создания прототипа заложены 4 этапа, необходимые для того, чтобы сеть приобрела профессиональную направленность.

Далее под модулем будем понимать разработанный фрагмент программы, который выполняет отведенную ему функцию. В настоящее время модуль осуществляет только выборку полезных данных: слова и частоту их использования, необходимых для обучения нейросети, так как выборка ограничена (рис. 3). В дальнейшем на этом этапе получаемые сведения будут использоваться для классификации и автоматического отбора текста в базу.

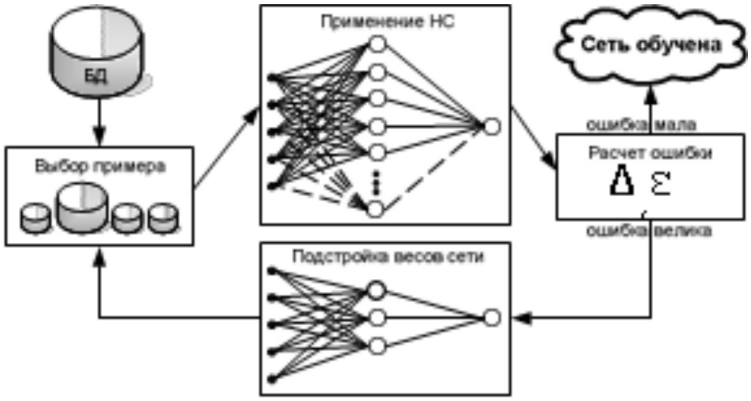


*Рис. 2. Иллюстрация процесса обучения*

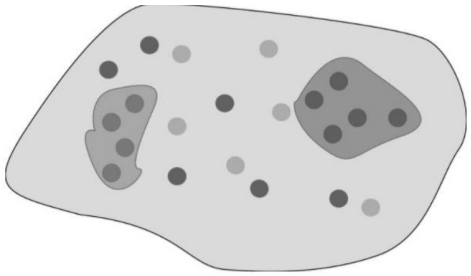

*Рис. 3. Иллюстрация процесса выборки полезных данных*

При использовании сети для решения практических задач необходимо иметь заранее заданную ее конфигурацию и держать постоянно сохраненной. За это отвечает модуль конфигурации, позволяющий создать и записать в память слепок нейросети. Именно на этом этапе происходит её дальнейшее обучение и использование. Задаются оптимальные значения последующего ее развития. Для реализации будет применена структура Нейронной сети прямого распространения (feed forward neural networks или FFNN), перцептроны (perceptrons) очень прямолинейны и передают информацию от входа к выходу (рис. 4).

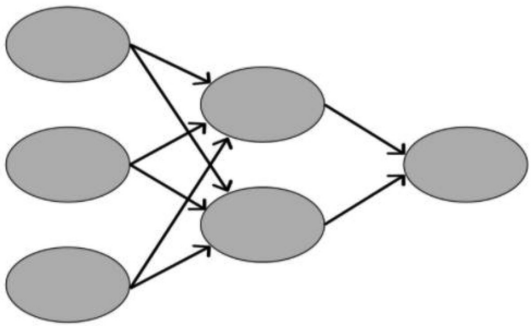

*Рис. 4. Иллюстрация структуры нейронной сети прямого распространения*

На основе собранной информации в этапе тренировки нейронная сеть получает множество других текстов и ожидаемый результат. Далее модуль вычисляет ошибку и делает переоценку своих решений. Вычисление ошибок происходит при помощи сигмовидной функции их вычислений, а переоценка решений – методом обратного распространения ошибки. Такой подход позволяет впоследствии сети делать более точные предсказания.

Этап использования предполагает рабочее применение нейросети. Задаются произвольные входные данные, которые сеть классифицирует, основываясь на ранее полученных данных. Сведения, которые приходят на вход, анализируются при помощи модуля для сбора дата сета.

Сформированная база данных работ сотрудников и обучающихся становится результатом с интеллектуальным управлением. На следующем цикле она употребляется в качестве дата сета нейронной сети, используя сквозную технологию не только в чтении статей.

## Заключение

Внедрение предлагаемой системы в университете позволит облегчить работу сотрудников по проверке курсовых и выпускных квалификационных работ обучающихся и по написанию собственных научных работ. Упростит процедуру взаимодействия преподавателей различных подразделений вуза между собой и объединения их в новые научные коллективы, охватывающие различные области науки. Позволит положительно влиять на публицистическую активность и качество научных статей. Поднимет корпоративную культуру, так как авторы будут видеть схожие работы.

### Список литературы

1. Галушкин А.И. Нейрокомпьютеры: учебное пособие. М.: Альянс, 2014. 528 с.

2. Кугаевских А.В. Модели и методы распознавания иероглифических текстов на примере древнеегипетского языка. автореф. дис. … канд. техн. наук. Тюмень, 2012. 21 с.

3. Ясницкий Л.Н. Интеллектуальные системы: учебник. М.: Лаборатория знаний, 2016. 224 с.

4. Васильев А.Н., Тархов Д.А. Принципы и техника нейросетевого моделирования. М.: Огни, 2015. 613 с.

5. Гелиг А.Х., Матвеев А.С. Введение в математическую теорию обучаемых распознающих систем и нейронных сетей: учебное пособие. М.: Издательство СПбГУ, 2014. 224 с.

6. Блащик Я., Блиновская К., Вуйчик Г.М., Гурецкий А., Дзедзицкая-Василевская М., Дурка П., Дух В., Жигеревич Я., Каминьский В.А., Лазаревич М.Т., Пшеласковский А., Склинда К., Стшелецкий М., Тадусевич Р., Хесс Г., Чишек Б., Шмяловская М. Основы нейрокибернетики. М.: РГГУ, 2015. 372 с.

7. Тархов Д.А. Нейросетевые модели и алгоритмы. Справочник. М.: Радиотехника, 2014. 920 с.

8. Рассел С., Норвиг П. Искусственный интеллект. Современный подход / Пер. с англ. М.: Вильямс, 2015. 1408 с.

9. Редько В.Г. Эволюция, нейронные сети, интеллект: Модели и концепции эволюционной кибернетики. М.: ИЛ, 2017. 224 с.

10. Иоголевич Н.И., Лободенко Е.И. Академическая недобросовестность студентов технического вуза: масштабы проблемы и пути решения // Педагогика. Вопросы теории и практики. Изд-во Грамота. 2020. Т. 5. № 1. С. 99–106.

11. Каляев И.А., Гайдук А.Р., Капустян С.Г. Модели и алгоритмы коллективного управления в группах роботов. М.: Физматлит, 2009. 280 с.

12. Игнатьев Н.А. Извлечение явных знаний из разнотипных данных с помощью нейронных сетей // Вычисл. технологии. 2003. Т. 8. № 2. С. 69–73.

13. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности: Справочное издание. М.: Финансы и статистика, 1989. 247 с.

14. Хокинс Дж., Блейксли С. Об интеллекте. М.: Изд. дом «Вильямс», 2007. 380 с.