УДК 004.02

ПРОБЛЕМЫ И ЗАДАЧИ ИДЕНТИФИКАЦИИ АВТОРА ТЕКСТА

Труханович И.А., Парамонов А.И.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Парамонов А.И. – кандидат технических наук

Аннотация. В работе приводится описание проблемы идентификации автора текста. Рассматриваются основные задачи, которые возникают в процессе исследования текста, и указаны существующие подходы к их решению. Сформулирована задача дальнейшая задача исследования вопроса определения авторства.

Ключевые слова. Идентификация автора текста, классификация текстов, авторский инвариант.

Идентификация автора текста заключается в подтверждении уверенности, что исследуемый текст составлен конкретным автором. Этот процесс основан на изучении других произведений этого автора и сопоставлении его «почерка». Решение задачи определения авторства может быть полезно для выявления наиболее вероятных авторов, а также поиска доказательств в поддержку предположений об авторстве. Его можно применять во многих задачах, таких как определение авторского права, обнаружение плагиата, анализ киберпреступлений и классификация сообщений (отнесение их к заданной категории, например, к спаму) [1]. Проблему идентификации автора текста можно считать междисциплинарной, поскольку для ее решения используются методы и модели когнитивной психологии, лингвистики, искусственного интеллекта и др. Ключевыми вопросами в ней можно обозначить поиск информации и обработку естественно-языковых текстов.

В цифровом пространстве известны нарушения авторских и смежных прав текста, которые выражены в использовании текста другого автора для материальной выгоды или наоборот – попыткой выдать авторство созданного текста за авторство другого человека. Эффективность защиты интеллектуальной собственности в цифровом пространстве определяется способностью противостоять таким нарушениям и угрозам их возникновения. Методы определения авторства позволяют пресечь такие нарушения и установить личность создателя текста [2].

За последние десятилетия мы наблюдаем постоянно растущий объем научных результатов. Такие области исследований, как библиометрия, а также наукометрия, нацелены на измерение и количественную оценку научных результатов. Этот постоянный рост объема научных публикаций создает серьезные проблемы, ведущие к включению идей из области идентификации авторов, для улучшения процессов измерения и анализа. Существуют предложения добавить понятие авторской атрибуции в предварительную обработку анализа научных публикаций, а также работы по отнесению сегментов статей к отдельным авторам. Таким образом, существует дискуссию о неявном определении научного авторства. Например, во многих областях науки предполагается, что первый автор выполнил большую часть (написания) работы, а последний автор внес свои идеи, будучи главой исследовательской группы. Таким образом, эффективность классификации представляет собой количественную оценку информации, содержащейся в стилометрии научной статьи, о числе авторов, участвовавших в ее написании [3].

Сегодня общение по электронной почте или с помощью мессенджеров стало повседневной нормой. Преступники злоупотребляют этими каналами текстовой информации для различных незаконных целей, таких как спам, незаконный оборот наркотиков, фишинг, клевета, травля и др. Некоторые киберпреступления, такие как кражи личных данных, интернет-мошенничество предусматривают раскрытие истинную личность автора, чтобы виновные могли быть наказаны в суде путем сбора убедительных доказательств против них. Криминалистический анализ может сыграть здесь решающую роль, позволяя судебному следователю собирать доказательства путем изучения подозрительных учетных записей электронной почты и социальных сетей. В этом контексте автоматическая идентификация авторства может помочь судебному следователю в расследовании киберпреступлений с использованием маркеров стиля, особенностей структуры и содержания.

В качестве «почерка» автора принято использовать авторский инвариант. Это количественная характеристика литературных текстов или некий параметр, который однозначно характеризует своим поведением произведения одного автора или небольшого числа "близких авторов", и принимает существенно разные значения для произведений разных групп авторов [4].

Решение данной проблемы можно разбить на следующие задачи:

1. Выбор модели представления текстов.
2. Выбор признаков для формирования авторского инварианта.
3. Выбор метода классификации с подходящими настройками.
4. Определение автора текста из множества предполагаемых авторов.

Каждая из задач на сегодня имеет уже ряд предполагаемых решений, которые используют различные подходы и имеют определенные результаты.

Для представления текстов в информационной системе можно использовать модель “мешок слов”, N-граммную и сглаживающую модели. Модель “мешок слов” представляется как неупорядоченная коллекция всех слов (или признаков слов), из которых состоит текст. В N-граммной модели текст понимается как последовательность цепочек из n элементов. Сглаживающие модели помогают справиться с проблемной разреженных данных в N-граммном представлении с помощью специальных техник сглаживания. На практике для модели “мешок слов” используются экспертные словари ограниченного объёма. В отличие N-грамм, данная модель не учитывает порядок компонентов текста. Кроме того, несмотря на некоторую относительную примитивность использования N-грамм для отображения авторского стиля и текста, на практике они являются эффективной описательной моделью, применимой ко многим языкам, а также в других областях. В частности, N-граммы очень хорошо показываются себя при решении задач криминалистики. Из недостатков можно отметить, что такое представление не может использоваться для описания разнесенных структур.

В качестве признаков, использующиеся для формирования авторского варианта, используются лексические, морфологические, синтаксические, структурные, контентно-специфические и другие.

К лексическим признакам относятся характеристики символов и слов, лексикон автора, маркеры стилей, распределение длин слов и др. Лексические признаки являются достаточно часто используемыми, поскольку одним из самых несложных методов засвидетельствовать или опровергнуть авторство текста представляется применение признаков отличительных особенностей словаря автора. Очевидно, что применение конкретных лексических единиц автором может быть весомым признаком его индивидуальности. Человек, обладающий состоятельным лексическим запасом, высказывает свои мысли, в большинстве случаев, больше емкими словами и фразами, преимущественно релевантными описываемой ситуации, его речь более определенна и ярка. Люди с незначительным лексиконом вынуждены довольствоваться использованием одних и тех же слов, поэтому их речь является более примитивной. Применение лексических признаков сталкивается с такими трудностями, как то, что характерные особенности у текста и автора могут отсутствовать. Кроме того, в случае наличия выраженные особенности, существует большая вероятность их подмены. Также определение отличительных черт авторского языка сопряжено с большой долей субъективизма.

К морфологическим признакам относят исследования расположения частей речи и полного набора грамматических классов. Недостаток подходов, основанных на словаре, заключается и в том, что они не учитывают словоизменение. Автор, персональной особенностью которого представляется использование определенного слова, как правило, применяет его различные формы. В большинстве случаев, образование новых слов происходит с помощью морфем, основа же остается постоянной.

Признаки, связанные со способами образования и использования словосочетаний и предложений, относят к группе синтаксических признаков. Такие признаки могут указывать на авторство, поскольку синтаксические конструкции образуются бессознательно во время создания текста, аналогично происходящему в устной речи, и не контролируются автором преднамеренно. Применение синтаксических признаков сталкивается с такими трудностями, как определение сложности, распространенности и состава предложения. Наиболее существенной проблемой является само проведение синтаксического анализа и возникающие в процессе его выполнения ошибки. С проблемами приходится встречаться начиная с определения границ предложения из-за неоднозначностей, существующих в языке. Неполнота применяемого морфологического словаря, омонимия и отсутствие эффективных алгоритмов морфологического анализа приводят к ряду сопутствующих сложностей. В конечном итоге становится невозможным определение устойчивых синтаксических конструкций.

Структурные признаки – это форматирование, цветовая гамма, оформление элементов текста и т.д. Структурные признаки могут указывать на авторство, поскольку нередко носят в себе отпечаток авторский представлений. К важным признакам относятся фрагменты текста (абзацы, главы), заголовки и стиль, цитирования различных источников. Характеристики для печатных документов могут включать пробелы, табуляции, отступы. Применение структурных признаков сталкивается с такими трудностями, как: последствия коррекции разметки сторонними людьми перед публикацией; последствия конвертирования документа из одного формата в другой; степень знания авторов методов форматирования.

К контентно-специфическим признакам относят тематические ключевые слова и фразы. Главной идеей, на которой базируются методы на основе таких слов и фраз, состоит в том, что если слово часто встречается в текстах одного класса, но редко — в текстах другого, то оно, скорее всего, имеет более весомое значение для разделения двух классов, чем слово, встречающееся в малом количестве текстов, но во многих классах. Применение контентно-специфических признаков сталкивается с такой трудностью, как эффективность только в рамках определенной тематики.

Характеристики текстового документа могут быть сведены в схему (рисунок 1).

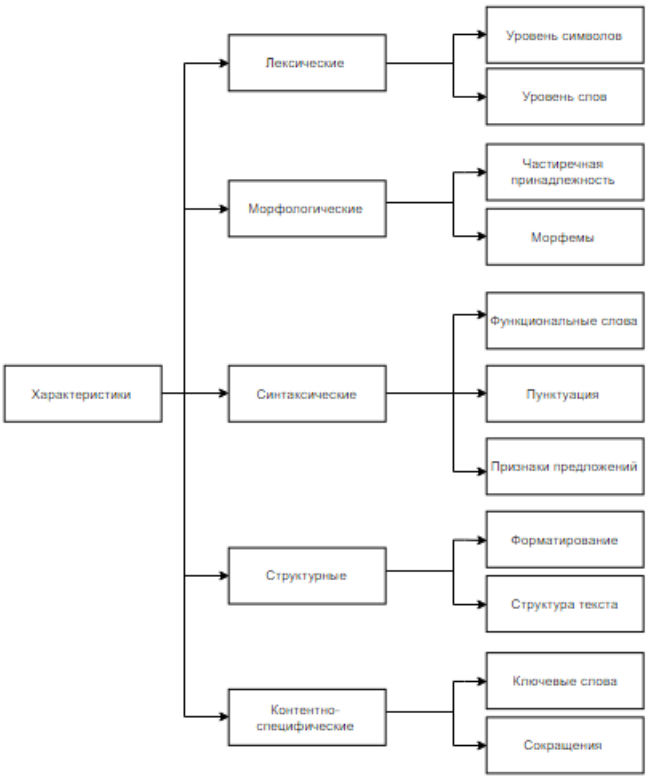

Рисунок 1 – Характеристики текстового документа

Эксперименты на различных объемах и жанрах показывают, что в настоящее время нейронные сети и метод опорных векторов при должном настройке и выборе входных параметров, являются наиболее перспективными. Но и у них есть свои недостатки. Нейронные сети требуют времени на обучение и повышенного внимания при работе с большим признаковым пространством. Временные затраты на подбор топологи сети и обучение можно сократить с помощью методов автоматического подбора топологии. Однако в итоге это сказывается на точности результатов. В метод опорных векторов отсутствуют такие проблемы, но при этом он чувствителен к шумам в исходных данных [5].

Исследования в рамках данной работы нацелены на рассмотрение эффективности, производительности и применимости означенных методов для идентификации авторства текстов. Уверенность в авторском инварианте и соблюдение требований к точности выходных данных в конечном счете предопределяет возможность определение авторства текста и формирование окончательного решения об самом авторе.

Список использованных источников:

1. Орлов, Ю.Н. Методы статистического анализа литературных текстов / Ю.Н. Орлов, К.П. Осминин. – М. : URSS, 2012. – 326 с.

2. Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks [Electronic resource] : MDPI. – Mode of access: https://www.mdpi.com/1999-5903/13/1/3/htm. – Date of access: 02.04.2021.

3. Authorship identification of documents with high content similarity [Electronic resource] : Springer. – Mode of access: https://link.springer.com/article/10.1007/s11192-018-2661-6. – Date of access: 03.04.2021.

4. Карта слов [Электронный ресурс]. – Режим доступа: https://kartaslov.ru/карта-знаний/Авторский+инвариант. – Дата доступа: 03.04.2021.

5. Wallace: Author Detection via Recurrent Neural Networks [Electronic resource] : Stanford University. – Mode of access: https://cs224d.stanford.edu/reports/YaoLeon.pdf. – Date of access: 03.04.2

UDC 004.02

ABOUT THE PROBLEM OF TEXT AUTHOR IDENTIFICATION

Trukhanovich I.A., Paramonov A.I.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Paramonov A.I. – PhD of Computer Sciences

Annotation. The paper describes the problem of text author identification. The main tasks that arise in the process of studying the text are considered, and the existing approaches to their solution are indicated. The task of further investigation of the problem is formulated.

Keywords. Text author identification, texts classification, authorial invariant.