МЕТОДЫ АНАЛИЗА ЦИФРОВОГО ТЕКСТА ДЛЯ ИДЕНТИФИКАЦИИ ЕГО АВТОРА

Парамонов А. И., Труханович И. А.

Белорусский государственный университет информатики и радиоэлектроники, Минск, Беларусь

Идентификация авторства использует методы поиска информации и обработки естественно-языковых текстов с целью определения личности, кем был составлен исследуемый фрагмент текста. Этот процесс основан на изучении других документов этого автора и сопоставлении его «почерка». Решение задачи определения авторства может быть применено во многих сферах, таких как авторское право, анти-плагиат, анализ киберпреступлений, классификация сообщений и др. Методы идентификации авторства текста сегодня представлены действенными инструментами в криминалистике для разрешения вопросов о спорном авторстве, построение портрета автора, стиля и т. д.

Современные подходы к вопросу определения авторства используют разнообразный математический аппарат, среди них наиболее известны следующие методы: *критерий Стьюдента, меры расстояния, Байесовский классификатор, метод ближайших соседей, метод опорных векторов* и др.

С помощью критерия Стьюдента проверка авторства текста может производиться на основе средних значений некоторых характеристик [1]. Например, по среднему количеству букв в словах или по средней длине предложения. Первоначально производится объединение текстов автора в единый. После этого тексты разделяются на одинаковые выборки и получают для каждого текста три характеристики: число выборок n_1, значение среднего m_1 и стандартное отклонение σ_1. То же самое проводится с текстом, полученным в результате объединения текстов, принадлежащих автору (n_2, m_2, σ_2). Затем для каждого текста рассчитывается значение t-характеристики критерия. Среди текстов, принадлежащих автору, определяется тот, значение критерия t которого максимально. Метод достаточно нетребователен к ресурсам, но требователен к объёму текста и обладает относительно низкой точностью, может использоваться при определении авторов журнальных статей.

Проверка авторства с помощью евклидового расстояния между частотными словарями выполняется на основании минимальности расстояние между частотными словарями для каждого исследуемого документа/текста [2]. Частотный словарь автора составляется на основе его документов/текстов. Затем выполняется нормализация словарей, при этом уникальные слова, которые есть только в одном из словарей, пропускаются. Метод относительно нетребовательный по ресурсам, но достаточно хорошо проявляет себя на текстах одного жанра одинаковых предметных областей разного большого объёма.

Метод ближайших соседей может использоваться для определения авторства на основе сходства значений атрибутов текста [3]. В n-мерном пространстве атрибутов все тексты соотносятся к классу определённого автора (либо неизвестного). Метод является достаточно ресурсоёмким, но не очень требователен к объёмам текста и обладает приемлемой точностью, может применяться в классификации текстов коротких электронных писем.

Модели на основе машины опорных векторов достаточно близки к перцептронам и могут быть использованы в классификации текстов. Метод является достаточно ресурсоёмким, но на сегодняшний день является одним из самых точных, может применяться на текстах различных жанров и объектов.

Кроме прочего, в вопросе авторства текстов можно использовать нейронные сети [4]. Они могут обладать весьма высокой точностью и быть применимы для разных предметных областей, но требуют значительных затрат на обучение.

В последнее время все чаще в обработке текстов используются генетические алгоритмы [5]. Подход на основе генетических алгоритмов является на сегодня самым гибким, поскольку в нём может подбираться один из наиболее эффективных методов для каждой конкретной задачи. Вместе с тем его результаты и время работы не всегда могут окупаться в силу самой природы данного подхода. Ведь сначала идёт подбор множества правил, которые описывают признаки текста. Затем правила проверяются на текстах, авторство которых известно, для каждого оценивается значение функции приспособленности. Таким образом происходит процесс отбора. Прошедшие отбор правила подвергаются изменению (мутации), и затем процесс отбора повторяется достаточно много раз, пока не отберутся авторские правила.

Анализ существующих подходов для определения авторства в разрезе различных характеристик даёт возможность оценить состояние текущих возможностей в данной области. Методы были проверены в рамках различного рода задач, некоторые из которых были весьма специфическими. В целом можно сделать выводы, что простейшие методы (критерий Стьюдента, Байесовский классификатор) обладают весьма низкой точностью, применяются, как правило, только в комплексе с другими или с модификациями. Однако, достаточно тривиальный метод расстояний может успешно использоваться при определении авторства художественных текстов одного жанра, но в остальных случаях он, как правило, недостоверен. Вполне приемлемые результаты дает метод ближайших соседей, который применим в области художественных текстов, а также и при сравнении журнальных статей, хоть и является затратным относительно выбранных признаков. Самыми же достоверными и точными можно назвать метод опорных векторов и нейронных сетей. Они дают высокую степень точности на текстах разных областей даже достаточно небольшого размера. Поэтому их можно использовать в достаточно широком диапазоне задач: от анализа диалогов в сети Интернет до авторства больших научных работ.

Литература

1. t-критерий Стьюдента [Электронный ресурс] / проект Математика. – Режим доступа: https://math.wikia.org/ru/wiki/T-критерий_Стьюдента. – Дата доступа: 25.03.2021.

2. Шумская А. О. Оценка эффективности метрик расстояния Евклида и расстояния Махаланобиса в задачах идентификации происхождения текста / А. О. Шумская // Доклады ТУСУР. – 2013. – № 3(29). – С. 141–145.

3. Агеев, М.С. Метод эффективного расчета матрицы ближайших соседей для полнотекстовых документов. / М.С. Агеев, Б.В. Добров // Вестник Санкт-Петербургского университета. 2011. №3 - с.72-84.

4. Федоров А.А. Анализ авторства текста полносвязной нейронной сетью в Keras [Электронный ресурс] – Режим доступа: https://www.bizkit.ru/2019/10/23/14754. – Дата доступа: 25.03.2021.

5. Бодянский Е.В., Волкова В.В., Коваль К.В. Автоматическая кластеризация текстовых документов на основе генетического алгоритма с искусственным отбором // Радіоелектроніка, інформатика, управління. 2009. №2 (21). – С. 91–96.