ОБЗОР РЕШЕНИЙ ЗАДАЧИ ИДЕНТИФИКАЦИИ АВТОРА ТЕКСТА

Труханович И.А., Кунцевич В.С.

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

Парамонов А.И. – канд. техн. наук, доцент

Идентификация автора текста актуальная задача в области обработки естественно-языковых текстов. Описаны основные задачи, которые возникают в процессе исследования текста, и указаны существующие методы их решения.

Высокая значимость идентификации автора печатного текста обуславливается быстром ростом объёма информации и стала одним из ключевых методов обработки и организации текстовых данных. Исследование идентификации авторов полезно для выявления наиболее вероятных авторов и поиска доказательств, подтверждающих вывод.

Таким образом, метод определения автора позволяет идентифицировать наиболее возможного автора из группы кандидатов в авторы научных статей, новостей, электронных писем и т.д. Этот процесс основан на изучении набора документов этого автора и сопоставлении его «почерка». В качестве «почерка» автора принято использовать авторский инвариант. Это количественная характеристика литературных текстов или некий параметр, который однозначно характеризует своим значением документы одного автора или некоторой группы «близких» авторов, и принимает существенно разные значения для произведений разных групп авторов.

В настоящее время замечается обостренный интерес к количественным технологиям разбора текстовой информации на основе слабо регулируемых автором черт текста, общих для всех авторов. Но общепризнанного определения о том, какой комплект черт даёт лучший результат, не существует. Достаточно мало внимания уделено идентификации автора с использованием комплексных черт текста.

Решение данной проблемы можно разбить на следующие задачи:

— Выбор модели представления текстов.
— Выбор признаков для формирования авторского инварианта.
— Выбор метода классификации с подходящими настройками.
— 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.

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

В качестве методов классификации применяются самые различные подходы: критерий Стьюдента, меры расстояния, байесовский классификатор, метод ближайших соседей, генетические алгоритмы, нейронные сети и метод опорных векторов.

Самыми 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. Тем не менее, они являются и ресурсозатратными. Что является существенными недостатками. Также нейронные сети требуют повышенного внимания при работе с большим признаковым пространством. Затраты времени на подбор топологи сети и обучение можно сократить с помощью методов автоматического подбора топологии, что в итоге сказывается на точности результатов. В метод опорных векторов отсутствуют такие проблемы, однако он чувствителен к шумам в исходных данных.

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