РАЗДЕЛ 7. КРАТКИЕ СООБЩЕНИЯ

УДК: 519.25; 004.8

ИДЕНТИФИЦИРУЮЩИЕ ПРИЗНАКИ ТЕКСТОВЫХ СООБЩЕНИЙ ПРИ УСТАНОВЛЕНИИ АВТОРА

А.О. Шумская

В работе приводятся основные подходы к установлению авторства, выделяется роль исследуемых при этом идентифицирующих признаков текста. Приводятся основные свойства идентифицирующих признаков текстов, используемых при атрибуции.

Ключевые слова: текст; авторство; идентифицирующий признак; атрибуция.

Введение

Вопросы, связанные с атрибуцией (установлением авторства) различных текстов, являются важными в области информационной безопасности, особенно на фоне увеличения объема текстовых массивов и появления новых способов для их распространения.

В числе таких задач подтверждение или исключение авторства того или иного лица, верификация авторства, определение метода создания текста, а также другие вопросы и задачи, которые так или иначе требуют определить, кем был создан текст. Разнообразие поставленных в связи с этим задач также обусловлено и различными формами текстов, имеющих различия по объему, стилю, происхождению, назначению и пр.

Идентификация авторства определяется как процесс установления автора по совокупности общих и частных признаков текста, составляющих авторский стиль [1].

Методы атрибуции текстов

В практике установления авторства долгое время преобладали субъективные методики, в соответствии с которыми отбирались внешние детали авторского стиля, такие как любимые слова, термины, выражения. Применение математико-статистических методов было начато в конце XIX века. В период с конца 70-х годов XX века отмечается повсеместная автоматизация обработки текстов и расчета их численных характеристик [2].

Современные методы атрибуции текстов основываются на статистическом анализе, либо на машинном обучении.

Методы статистического анализа берут за основу тот факт, что стиль автора можно определить по какому-то определенному параметру или набору таких параметров – так называемый авторский инвариант. Так как с увеличением объема текста параметры, его составляющие, становятся устойчивыми с вероятностной точки зрения. Это позволяет устанавливать авторство по характеристикам текста [3].

Суть статистических методов идентификации авторства заключается в определении критической границы для некоторого авторства, с которой сравнивается текущее значение. И, в зависимости от положения на числовой оси этого значения относительного критической границы, делается заключение о том, что текст с высокой вероятности принадлежит автору, либо, напротив, с высокой вероятностью не принадлежит автору.

Некоторые наиболее распространенные методы идентификации авторства, основанные на статистическом анализе:

- Критерий Стьюдента,
- Хи-квадрат,
- Критерий Колмогорова-Смирнова,
- Марковские цепи,
- Энтропийный подход и др.

Методы, основанные на машинном обучении, заменяют мнение эксперта результатами работы систем принятия решений. Если при статистическом анализе эксперт отбирает ряд характеристик, составляющих инвариант, то данные механизмы позволяют выделить инвариант автора, отличающий его от остальных [4].

Однако выработанные правила не всегда очевидны для человека, так как такие механизмы работаю по принципу «черного ящика» [4].

Некоторые наиболее распространенные методы идентификации авторства, основанные на машинном обучении:

- Нейронные сети,
- Метод опорных векторов,

- Генетические алгоритмы,
- Деревья решений и др.

Выбор идентифицирующих признаков

Методы атрибуции в общем случае исследуют текстовое произведение на пунктуационном, орфографическом, синтаксическом, стилистическом и лексико-фразеологическом уровнях.

Под авторским инвариантом понимается количественная характеристика текстов, которая однозначно характеризует своим поведением произведения одного автора или небольшого числа "близких авторов", и принимает существенно различные значения для произведений разных групп авторов [1]. Исследования в данной области показывают, что, чаще всего, авторский инвариант обусловлен тремя уровнями: синтаксическим (особенностями построения предложений), стилистическим (характерные речевые приемы) и лексико-фразеологическим (словарный запас автора).

Однако, многообразие грамматических структур, участвующих в формировании текстов, сильно затрудняет поиски таких инвариантов. Вычислительные эксперименты показывают, что обнаружение числовых характеристик, различающих разных авторов, - сложная задача. Дело в том, что при создании автором некоторого текстового произведения существенную роль играют не только подсознательные, но и сознательные факторы [5].

В работах, посвященных атрибуции текстовых форм [1], выделяются следующие признаки текстовых характеристик, составляющих авторский инвариант.

1) Массовость, под которой понимается свойство параметра слабо контролироваться автором на сознательном уровне;

2) Устойчивость, под которой понимается сохранение значения параметра в некотором диапазоне для одного автора;

3) Различающая способность, под которой понимается свойство текстовой характеристики принимать существенно различные значения для разных авторов. Существенное различие здесь понимается как превышение уровня колебания значения, учтенного выше как диапазон разброса значений для одного автора.

Выбор идентифицирующих признаков, которые бы гарантированно разделяли двух любых авторов, практически невозможен. Поэтому на практике считается достаточным, чтобы параметр позволял уверенно различать разные группы авторов, то есть существовало достаточно большое количество групп авторов, для которых средние значения параметра значительно различаются. Параметр в таком случае не поможет различить тексты авторов из одной группы, но позволит различать тексты авторов, попавших в разные группы. Различать тексты авторов одной группы можно за счет использования одновременно достаточно большого вектора различных по характеру параметров.

Заключение

Важнейшим элементом процесса атрибуции текста является определение вектора идентифицирующих признаков. Для успешной обработки текста признаки должны обладать свойствами массовости, устойчивости и различающей способности.

Определение вектора признаков в том или ином случае зависит от уровня исследования текста, определяется перечисленными свойствами признаков, а также может зависеть от специфики поставленной задачи.

СПИСОК ЛИТЕРАТУРЫ

1. Романов, А.С. Разработа и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста / А.С. Романов, А.А. Шелупанов, Р.В. Мещеряков. – Томск : В-Спектр, 2011. – 188 с.
2. Романов, А.С. Состояние проблемы распознавания и идентификации автора текста [Электронный ресурс] // Информационная безопасность. – Электрон. текст. дан. – [Б.м.], 2010-. – Режим доступа: http://inf-bez.ru/?p=813 (дата обращения: 18.03.2013).
3. Мещеряков, Р.В. Диалог как основа построения речевых систем/ Р.В. Мещеряков, В.П. Бондаренко // Кибернетика и системный анализ. 2008. № 2. С. 30-41.
4. Романов, А.С. Методика идентификации автора текста на основе аппарата опорных векторов [Электронный ресурс]/ А.С. Романов, // Доклады ТУСУРа. – 2009. – № 1 (19), ч. 2. – С. 36-42. – Электрон. версия печатн. публ. – Режим доступа:: http://www.tusur.ru/filearchive/reports-magazine/2009-1-2/36-42.pdf (дата обращения: 08.04.2013)
5. Родионова, Е.С. Методы атрибуции художественных текстов/ Е.С. Родионова// Структурная и прикладная лингвистика : межвуз. сб. / под ред. А.С. Герда. – СПб. : Изд-во СПб. гос. ун-та, 2008. – Вып. 7. – С. 118-127.

***Шумская А.О.**, инженер каф. КИБЭВС ТУСУР*