Ю. Н. Орлов, К. П. Осминин

Определение жанра и автора литературного произведения статистическими методами

В настоящей статье излагается метод классификации текстов на основе анализа статистических закономерностей буквенных распределений, т. е. вероятностей встречаемости букв и буквосочетаний. Подробно рассматривается задача кластеризации литературных произведений по определенным жанрам, а также вопрос определения авторства произведения. При этом решение должно быть найдено без вторжения в область литературы, т. е. без анализа синтаксиса, литературных приемов и схем взаимодействий персонажей.

Введение

По-видимому, впервые статистический анализ был применен к вопросу авторства литературного произведения почти сто лет назад А. А. Марковым [1]. Он предположил, что текст представляет собой случайную цепочку из гласных и согласных букв, связанных между собой определенными вероятностями перехода. Тогда авторство может быть установлено путем сравнения соответствующих вероятностей, которые предполагаются постоянными для каждого автора. Ограниченность метода состоит в том, что эти вероятности существенно зависят от объема текста, по которому они рассчитываются, и эволюционируют на протяжении всего произведения, так что погрешность метода оказывается слишком велика.

Тем не менее, хотя литературное произведение не является реализацией Марковского процесса, существуют разные модификации этого метода, поскольку он легко реализуем на практике. Интересным примером развития рассматриваемой методики стала работа Д. В. Хмелева [2], где уточняющим инструментом служит функция максимального правдоподобия, в качестве которой выбрана информационная энтропия для парных буквосочетаний.

В большинстве существующих методик предполагается некоторая инвариантность авторской манеры письма, что приводит к поиску различных «авторских инвариантов». Это может быть доля гласных или согласных, информационная энтропия, распределение используемых слов по длине, переходные вероятности между парами букв, доля союзных слов (см., например, А. Т. Фоменко [3]) и иные функционалы от распределения текста по буквам и буквосочетаниям. К сожалению, к любому методу, основанному на статистике (в том числе и к применяемому в настоящей работе), можно подобрать контрпример. В частности, таковым является роман «Улисс» Джеймса Джойса, каждая из восемнадцати глав которого написана в разном стиле. Авторами проверено, что ни один из известных методов не дает удовлетворительного ответа (в том смысле, что все главы написаны одним и тем же человеком) более чем по трем парам глав из 153 пар. Разумеется, это не является «противозаконным»: автор имеет право изменить стиль, начинать все слова с буквы «о» и т. д. Кроме того, при увеличении числа сравниваемых произведений возникает неизбежное сближение инвариантов, так что, начиная с какого-то количества авторов, расстояние между инвариантами становится меньше, чем среднеквадратичное отклонение инварианта, которым обычно пренебрегают. Поэтому такая методика имеет принципиальные ограничения.

Отметим также еще один недостаток существующих в данной области работ на-