Математические методы атрибуции литературных текстов небольшого объема

Рогов А. А., Сидоров Ю. В., Суровцова Т. Г.
rogov@psu.karelia.ru
Петрозаводск, Петрозавоский государственный университет

Изучение литературных произведений с использованием математических методов имеет богатую историю, а появление компьютеров расширило возможности проведения вычислительных экспериментов. Основной целью работы является поиск методов, которые помогут «оценить» стиль литературного текста, выявить закономерности, присущие разным жанрам, авторам, произведениям. Для этого необходимо на едином тестовом материале исследовать методы на надежность и устойчивость.

Материал для исследования

Основой для проведения исследования стала электронная коллекция публицистических статей из петербургских журналов «Время», «Эпоха», «Современник», «Гражданин», и других текстов 60–70-х гг. XIX века в оригинальной орфографии дореволюционной России [2, 3], которая создается в Петрозаводском государственном университете, начиная с 1995 года, включающая синтаксический и морфологический разборы произведений. Богатый материал дает основу для анализа авторского стиля. Причем можно рассматривать не только легко рассчитываемые признаки (длина предложения, средняя длина слова и т. д.), но и более сложные, описывающие грамматику текста. Что позволяет более полно изучать произведения, размер которых не достаточно велик.

Методы и результаты

Был проведен поиск методов, описанных в литературе, которые использовались для атрибуции текстов [1, 4], и их проверка на нашем материале. Предложены модификации методов, а также разработаны собственные подходы [3] и их приложение к атрибуции текстов небольшого объема. Наиболее интересные, с нашей точки зрения, вошли в экспертную систему, входящую в программный комплекс «Статистические методы анализа литературных текстов» (ПК СМАЛТ). В частности для проведения анализа в экспертной системе предлагаются следующие группы методов:

— разбиение анализируемых текстов на однородные группы с близким набором грамматических признаков;
— проверка статистических гипотез об однородности распределения частотных характеристик текстов, таких как распределение частей ре-

чи на разных позициях предложения, индекс разнообразия лексики и т. д.;

— метод «сильного графа» для оценки парной связи грамматических классов.

Наборы признаков для анализа, которые можно получить на основе грамматических разборов произведений, очень разнообразны. Сложно определить те, которые описывают авторский стиль, поэтому предусмотрена возможность проводить исследования на наборе морфологических и синтаксических признаков, который определяет самостоятельно специалист-филолог. Выбираются и методы, которые должны быть применены. С использованием экспертной системы ПК СМАЛТ был проведен ряд исследований. Одно из них по поиску авторского инварианта — некоторых особенностей текста, неосознанное предпочтение которым отдает автор при создании своих произведений, описано ниже. Были выбраны публицистические произведения, объемом от 6 до 700 предложений. Проведем краткое описание методов и результатов.

Анализ частоты синтаксических конструкций. Параметры, выбранные для анализа синтаксического разбора, представляют собой относительную встречаемость определенной синтаксической конструкции в тексте (тип предложения, осложненность, наличие второстепенных членов и т. д.). Для получения групп произведений использовался алгоритм иерархического кластерного анализа (методы ближайшего и дальнего соседа с евклидовым расстоянием и расстоянием Чебышева) [3]. Группы объектов, которые получились в результате, не дали четкого разделения по авторам произведений, причем группировка менялась в зависимости от выбранной метрики и метода построения кластеров. Выбранный для анализа набор признаков оказался неустойчивым.

Метод «сильного графа» основан на определении пороговых значений [1], которые позволяют оставить устойчивые связи, отбрасывая более редкие, как менее значимые. Получаемые в результате графы сравниваются для обнаружения близости между текстами. Была рассмотрена связь между синтаксическими структурами текста, выделены классы предложений: простое односоставное, сложное бессоюзное предложение и т. д. Изучались переходы между классами, которые присутствуют в текстах, строились «сильные графы» с различными пороговыми значениями. В результате был сделан вывод о том, что структура графа в первую очередь зависит от длины рассматриваемого текста. С помощью данного метода не удалось получить достаточно устойчивые результаты, то есть даже при небольших изменениях пороговых параметров, матрица близости текстов сильно меняла свой вид.

Проверка статистических гипотез об однородности распределения частотных характеристик текстов. Были использованы методы, приведенные в исследованиях Г. Хетсо [5]. Эти методы основаны на проверке статистических гипотез о значимости различий рассчитываемых параметров для сравниваемых авторов. Из числа рассматриваемых текстов исключаются те, которые имеют значение критерия, попадающее в критическую область. Полученные результаты согласуются с результатами, полученными Г. Хетсо, о возможной принадлежности некоторых произведений Dubia (спорное авторство) перу Ф. М. Достоевского, не смотря на то, что мы работали с текстами в оригинальной орфографии XIX века и отличными морфологическими разборами текстов. Что говорит об устойчивости методов и о выявленных реальных отличиях или сходстве.

Заключение

Использование ПК СМАЛТ позволяет проводить проверку методов на разных наборах грамматических признаков. Так некоторые из методов оказались неустойчивыми для выбранных нами наборов признаков. Происходит поиск новых методов, которые могли бы работать с произведениями небольшого объема. В настоящее время разрабатывается комплексный подход к проводимым экспериментам, который учитывал бы результаты, показанные каждым из методов. Все полученные данные предложены для рассмотрения специалистам-филологам, которые исследуют творчество Ф. М. Достоевского.

Проект поддержан грантами РГНФ № 02-04-12015в, № 05-04-12418в.

Адрес проекта в Интернете: `http://smalt.karelia.ru`.

Литература

[1] *Бородкин Л. И., Милов Л. В., Морозова Л. Е.* К вопросу о формальном анализе авторских особенностей стиля в произведениях Древней Руси // Математические методы в историко-экономических и историко-культурных исследованиях. — Москва, 1977. — С. 298–326.

[2] *Захаров В. Н., Леонтьев А. А., Рогов А. А., Сидоров Ю. В.* Программная система поддержки атрибуции текстов статей Ф. М. Достоевского // Труды Петрозаводского государственного университета: Сер. Прикладная математика и информатика. Вып. 9. — Петрозаводск: ПетрГУ, 2000. — С. 180–189.

[3] *Rogov A. A., Sidorov Yu. Vl.* Statistical and Information-calculating Support of the Authorship Attribution of the Literary Works // 6th Int. Conf. Computer Data Analysis and Modeling: Robustness and Computer Intensive Methods, Vol. 2: K-S. — Minsk: BSU, 2001. — Pp. 187–192.

[4] *Хетсо Г.* Принадлежность Достоевскому: к вопросу об атрибуции Ф. М. Достоевскому анонимных статей в журналах Время и Эпоха // Oslo: Solum Forlag A. S., 1986. — 82 с.