УДК 681.301

К.т.н., доцент Россошинський Д. О., магістрант Звєрєв О. Ю.

**Національний технічний університету України
«Київський політехнічний інститут»**

АВТОМАТИЧНЕ ВИЗНАЧЕННЯ АВТОРСТВА ТЕКСТУ

Вступ

Сутність проблеми встановлення авторства літературних текстів (authorship attribution) - визначити, хто є автором твору за деякими характеристиками тексту [1]. Це буває потрібно при визначенні плагіату, знаходженні різних псевдонімів одного автора. Наприклад, відоме використання формальних методів при перевірці авторства роману М.А. Шолохова «Тихий Дон» [2], аналіз авторства тексту, який приписують Салтикову-Щедріну [3]. Трапляються й анонімні твори, авторство яких можна визначити, порівнявши їх з іншими текстами, автор яких відомий. Дослідження текстів може застосовуватися в літературознавстві, історіографії, криміналістиці, захисті авторського права та інших галузях.

Постановка задачі

Є деякий твір, дані про авторство якого невідомі або вимагають перевірки. Існує також скінчена множина творів, авторство яких відоме. Необхідно визначити, кому з перелічених авторів з найбільшою ймовірністю належить твір, або ж він значно відрізняється від усіх. При автоматичному визначенні авторства найбільш ймовірний претендент визначається за допомогою комп’ютерного аналізу тексту твору та порівняння його з іншими зразками творів.

Підходи до задачі визначення авторства

Існує ряд підходів до визначення авторства тексту [4]. Їх можна розділити за ступенем впливу людини на процес атрибуції.

- Неформальні методи (суб'єктивно-атрибутивна методика) - експерт проводить аналіз самостійно, без допомоги ЕОМ, покладаючись на свій досвід та інтуїцію.
- Формальні математичні та статистичні методи (метод накопичувальних сум, метод лінгвістичних спектрів, метод опорних слів,

метод ланцюгів Маркова, метод ущільнення даних) - аналіз тексту проводиться на ЕОМ за деяким алгоритмом, без участі людини.

- Методи штучного інтелекту (використання штучних нейронних мереж, генетичні алгоритми) - здатні «навчатись» на деяких вхідних текстах і далі самостійно приймати рішення про авторство текстів.
- Комбіновані математично-психолінгвістичні методи (метод семантичного диференціалу) - група експертів оцінює текст за деякою сукупністю шкал, потім дані, отримані в результаті опитування, проходять математичну обробку.

Прототип веб-служби, що використовує метод лінгвістичних спектрів

В [1] нами був розроблений прототип програми, що реалізує метод лінгвістичних спектрів, який був запропонований Морозовим у [5].

Суть методу лінгвістичних спектрів - відмінність стилів різних авторів на основі частоти використання ними в текстах службових слів: сполучників, прийменників, а також займенників та прислівників. Ці слова вживаються майже у всіх видах літератури і не залежать від змісту тексту. Один з різновидів методу розглядає головний прийменниковий спектр - кількість повторень в тексті прийменників «в», «на» і «с». Існують також сполучникові, займенникові та інші спектри.

Програма розроблена на мові PHP і є веб-сервісом, що дозволяє одночасно багатьом користувачам використовувати один примірник програми, а в подальшому - зберігати вхідні дані та результати в єдиній базі даних.

Програма аналізує введену частину тексту й будує діаграми (лінгвістичні спектри) на підставі підрахунку прийменників в тексті.

Отримані діаграми дозволяють наочно оцінити схожість стилю творів. Процес аналізу можна автоматизувати, використовуючи метод найменших квадратів і отримуючи чисельні значення, що характеризують схожість або несхожість текстів.

Набір макросів до пакету Microsoft Office для перевірки методу авторського інваріанту

Основним інструментом більшості авторів та редакторів сьогодні є текстовий редактор Microsoft Word. У ході дослідження було розроблено сукупність макросів, що об’єднують можливості Microsoft Word та табличного редактора Microsoft Excel та реалізують аналіз текстів методом опорних слів (авторського інваріанту).

Метод опорних слів [2] ґрунтується на підрахунку службових слів (14 сполучників, 38 прийменників і 17 часток - разом 55 службових слів) у фрагментах тексту фіксованої довжини. Тестування методу на великій кількості творів різних письменників дозволило авторам цього методу стверджувати, що ця характеристика є «авторським інваріантом» (величиною, що мало змінюється протягом усього періоду творчості письменника).

У процесі експерименту було виявлено, що метод працює недостатньо точно при описаному в [2] наборі службових слів. В табл.1 наведено результати аналізу повного тексту декількох творів одного колективу авторів та творів іншого автора:

Таблиця 1. Результати пошуку авторського інваріанту текстів

Автор	Твір	Процент службових слів	Всього слів	Службових слів
Аркадий и Борис Стругацкие	Обитаемый остров	15,6175	128644	20091
Аркадий и Борис Стругацкие	Трудно быть богом	14,6200	72421	10588
Аркадий и Борис Стругацкие	Пикник на обочине	16,3880	69020	11311
Борис Акунин	Азазель	15,7276	78022	12271

Як видно з таблиці, розбіжність між текстами одного автора за цим методом, що складає 1,77%, може перевищувати різницю між текстами творів різних авторів, яка в даному випадку складає лише 0,11%, тому за цим параметром розділити тексти різних авторів неможливо. Згідно отриманих результатів, метод потрібно модифікувати, змінивши вибір опорних слів та пристосувати його до автоматичного пошуку в текстах великого обсягу.

Автоматизована система документообігу редакції

Зараз нами розробляється система, що автоматизує обіг документів у редакції газети. Її складовою частиною буде модуль, що проводить аналіз авторства текстів. При цьому буде використовуватись комбінація декількох методів: авторського інваріанту, лінгвістичних спектрів та марковських ланцюгів. Вагові коефіцієнти будуть визначені за результатами експериментів над базою опублікованих та поданих до публікації статей, у якій зазначені автори та особи, що редагували тексти, а також відмітки літературного редактора чи психолога щодо суб'єктивних характеристик тексту.

Висновки

Існує багато методів визначення авторства тексту, більшість з них дозволяє повністю або частково автоматизувати процес.

У процесі дослідження були розроблені веб-сервіс та набір макросів, що реалізують аналіз тексту методами лінгвістичних спектрів та авторського інваріанту та використовують різні інформаційні технології. Найбільш перспективними для програмної реалізації є також методи семантичного диференціалу та метод ланцюгів Маркова. Схожі методи можуть використовуватися для аналізу творів живопису, скульптурі, музиці, достатньо лише підібрати відповідні набори шкал та критеріїв оцінювання.

Можлива реалізація програми, що реалізує ці методи як елемента комп'ютерної системи чи інтернет-порталу, присвяченого літературі або захисту авторських прав.

Література

1. *Зверев А.Ю.* Автоопределитель авторства// Мой компьютер, №42 (473), 2007, С. 40-42.

2. *Фоменко В.П., Фоменко Т.Г.* Авторский инвариант русских литературных текстов Предисловие А.Т. Фоменко //Фоменко А.Т. Новая хронология Греции: Античность в средневековье. Т.2. – М.: Изд-во МГУ, 1996. – С. 768-820.

3. *Батов В.И.* Существует ли формула авторства//Число и мысль. Сборник, выпуск 7 – М.:Знание, 1984. – С. 117-137

4. *Зверев А. Ю.* Анализ текста для определения его авторства // Интеллектуальный анализ информации ИАИ-2008. Сборник трудов. – К.: Просвіта, 2008. – С. 208-218.

5. *Морозов Н.А.* Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного известного автора. Стилеметрический этюд// Известия отд. русского языка и словесности Имп. акад. наук. 1915. Т.20, Кн.4.