**УДК 004.912**

**магістрант Глушаускайте І.В., к.т.н., ст. викл. Заболотня Т.М.**

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

# ОСОБЛИВОСТІ ЗАСТОСУВАННЯ КОМБІНОВАНОГО АЛГОРИТМУ АВТОМАТИЗОВАНОГО ВИЗНАЧЕННЯ АВТОРСТВА ТЕКСТІВ

## Abstract

**Irina Glushauskaite, student; Tetiana Zabolotnia, PhD**
***Features of using the combined algorithm for automatic text attribution***

*In this work the algorithm of automated text authorship determination is analyzed. The peculiar tasks for which the developed method suites better are named, and further description of each such task in detail is given.*

## Вступ

Задача визначення авторства тексту вже багато років є актуальною як в теоретичній лінгвістиці, так і в галузі прикладних задач. Визначення авторства застосовується у філологічних дисциплінах, в психології та теорії штучного інтелекту, в судовій практиці та криміналістиці. Впровадження алгоритмів визначення та порівняння авторських стилів актуальне для вирішення задач інформаційного пошуку, все більш поширюваних у зв'язку зі збільшенням обсягів інформації і розповсюдженням мережі Інтернет.

Завдяки повсюдному переходу від рукописного тексту до електронного виникає безліч можливостей для впровадження систем комп’ютеризованого визначення авторства.

## Постановка задачі

Для підвищення ефективності процедури визначення автора текстів за критеріями точності, повноти на F-міри створений комбінований алгоритм, що поєднує використання засобів класифікації та кластеризації текстів.

Мета даного дослідження полягає в пошуку та аналізі аспектів задачі визначення авторства, для розв'язання яких використання розробленого алгоритму є найбільш ефективним

**Комбінований алгоритм автоматизованого визначення авторства тексту**

Беручи до уваги результати дослідження існуючих методів визначення авторства, було створено новий метод, що врахував їх переваги і недоліки. Даний метод включає етапи щодо визначення автора тексту з набору відомих програмі авторів (використовуючи засоби класифікації), а також етапи з автоматичного додавання до класифікатора категорій для нових авторів. Також авторами було запропоновано алгоритм реалізації методу, що може бути виконаний відповідними програмними засобами.

Згідно створеного алгоритму спочатку тексти, що представляють вхідні дані, розсортовані по деяких категоріях, але надалі при додаванні нових текстів засоби кластеризації дозволяють об'єднувати ці тексти в нові категорії. Нововведені категорії відповідають групам текстів, авторів яких не було в первинному списку.

**Особливості застосування алгоритму**

Запропонований алгоритм може бути застосований до класів задач, для яких з певних причин використання існуючих засобів визначення авторства текстів є недоцільним. У даному розділі наведемо та проаналізуємо ці задачі. Для зручності під існуючими засобами визначення авторства текстів будемо розуміти програмні засоби класифікації та кластеризації, оскільки вони використовуються у переважній більшості випадків.

1. *Автоматична розмітка текстових корпусів* за критерієм авторства, приклад – велика електронна бібліотека. Існує текстовий корпус, що періодично поповнюється новими текстами; експертний аналіз не є придатним у даній ситуації через надвеликі часові витрати. Задача програми для розмітки – встановлювати відповідність між текстами та їх авторами для подальшого поліпшення пошуку в корпусі та впорядкування текстів. При використанні засобів класифікації існують наперед задані категорії текстів, нові тексти належать до існуючих категорій. Але, яким би повним не був початковий корпус, час від часу з’являються нові автори. Віднесення тексту нового автора до категорії текстів старого автора є помилкою та веде до некоректної розмітки корпуса, а некоректна розмітка може призвести до помилок у подальшому використанні корпуса. Створений алгоритм дозволяє присвоювати текстам нових авторів «невизначену» категорію та створювати нову категорію після того, як накопичиться певна кількість схожих між собою текстів з «невизначеною» категорією. Це є особливо ефективним для новим авторів, яким відповідає

багато текстів – наприклад, для авторів, що пишуть не романи, а оповідання.

2. Через можливість визначати автора тексту як «невизначеного», запропонований алгоритм є простим рішенням для визначення належності певного тексту відомим системі авторам. Отже, алгоритм можна застосовувати *для визначення плагіату*.

3. Крім визначення того, чи належить певний текст відомим програмі авторам, запропонований алгоритм дає змогу додавати цей текст до загальної бази текстів програми. При додаванні кількох текстів, схожих за стилем, формується нова категорія, приналежність до якої також може бути визначена алгоритмом. Отже, запропонований алгоритм може застосовуватися для виявлення не лише плагіату з текстів, що були у програми на початку роботи, а і з будь-яких текстів, що надійшли до програми після початку її роботи. Це суттєво зменшує залежність програми від початкового наповнення бази текстів. Через це алгоритм може бути застосований *у сфері освіти для визначення авторства рефератів, курсових* тощо навіть в умовах плагіату з нових робіт і частої зміни навчальних дисциплін.

4. Аналогічно більш точне, ніж для стандартних алгоритмів, *визначення авторства анонімних творів,* анонімних листів, підроблених документів. Програмна реалізація алгоритму визначає, чи належить документ відомому їй автору, проте при надходженні даних про нових авторів вона здатна видати результат з урахуванням цих даних. Ця властивість може бути застосована, наприклад, у криміналістиці – деякий зловмисник відправляє анонімні листи із загрозливим змістом, експертиза нездатна встановити автора першого листа, проте здатна встановити схожість цих листів, як тільки стане відомим авторство одного з листів – можна буде казати про аналогічне авторство для інших листів.

5. Застосування алгоритму для визначення унікальності текстів, а також для визначення автора анонімних текстів, може бути використано *для аналізу публікацій в мережі Інтернет*. Наприклад, алгоритм може використовуватись пошуковими системами для індексації та пошукової оптимізації Інтернет-сторінок.

6. Алгоритм, подібний до запропонованого, може бути застосований до низки *задач з галузі психолінгвістики*. Він буде доцільним для побудови точної класифікації за деякою психолінгвістичною категорією при недостатніх початкових даних. Наприклад, перед дослідником стоїть задача аналізу корпусу україномовних текстів, написаних людьми, для яких українська є іноземною. Дослідник володіє експериментальними даними – низкою текстів, що написані іспаномовними людьми, та текстів, що написані людьми німецькомовними. Комбіноване використання

методів класифікації та кластеризації стає у нагоді в цій задачі. Можна використовувати подібний алгоритм для категоризації текстів за відповіддю на якесь питання психолінгвістики, що містить невідому наперед кількість відповідей, у разі, коли непросто підібрати тексти для категорій, що відповідають кожній відповіді.

7. Через автоматизацію визначення нових категорій системи, що реалізують представлений алгоритм, більш *придатні до роботи без нагляду експерта* та роботи з рідко поновлюваними базами текстів.

**Висновки**

Комбінований алгоритм визначення авторства текстів, що поєднує в собі кроки щодо класифікації та кластеризації документів, дозволяє підвищити значення чисельних оцінок якості класифікації.

Крім цього, запропонований алгоритм може бути ефективно застосований до низки специфічних задач. У даній роботі описані особливості та переваги використання алгоритму для таких задач, наприклад – для розмітки текстових корпусів, для визначення приналежності тексту відомим авторам, для визначення плагіату в статтях та навчальних роботах, для пошукової оптимізації Інтернет-сторінок.

**Література**

1. *Агеев М. С.* Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов : дис. ... канд. физ.- мат. наук : — М., 2004. — 136 с.

2. *Романов А. С.* Методика и программный комплекс для идентификации автора неизвестного текста : дис. ... канд. техн. наук — Томск, 2009. — 149 с.

3. *Шевелев О. Г.* Разработка и исследование алгоритмов сравнения стилей текстовых произведений : дис. ... канд. техн. наук — Томск, 2006. — 176 с.

4. *Сидоров Ю. В.* Математическая и информационная поддержка методов обработки литературных текстов на основе формально-грамматических параметров : дис. ... канд. техн. наук — Петрозаводск, 2002. — 127 с.

5. *Васильев В. Г.* Методы автоматизированной обработки текстов – М. :ИПИ РАН, 2008. – 305 с.