УДК 004.056.5
В. Вівчарик
(Тернопільський національний технічний університет імені Івана Пулюя)

ОСОБЛИВОСТІ МЕТОДІВ АНАЛІЗУ ТЕКСТІВ ДЛЯ ІДЕНТИФІКАЦІЇ АВТОРСТВА ДОКУМЕНТУ

UDC 004.056.5
V. Vivcharyk
(Ternopil Ivan Puluj National Technical University, Ukraine)

PECULIARITIES OF TEXT DATA MINING METHODS FOR DOCUMENT AUTHORSHIP IDENTIFICATION

Як відомо, проблема аутентифікації повідомлення, тобто підтвердження автора повідомлення, є одним з наріжних каменів криптографії, яка традиційно розв'язується алгоритмами цифрового підпису. Проте існує кардинально інша проблема ідентифікації та встановлення автору документа шляхом аналізу контенту документу. Ця проблема є менш поширеною, проте не менш важливою, адже задача ідентифікації автора розв'язується для боротьби з плагіатом, встановлення авторства анонімних текстів, програмного коду, шкідливого програмного забезпечення, експертизи та встановлення особистості в криміналістиці, запобігання злочинів та багатьох інших застосувань.

Дослідження "авторського стилю" може здійснюватися на різних рівнях: пунктуаційному, орфографічному, синтаксичному, лексико-фразеологічному та стилістичному. Найбільший інтерес дослідників представляє аналіз трьох останніх рівнів. Існує доволі багато методів аналізу стилю. В цілому їх можна розділити на дві групи: експертні та формальні. Формальні методи базуються на алгоритмах статистичного аналізу, машинного навчання, нейронних мереж, Text mining та ін. Останнім часом зростає популярність методів вбудовування в мережі (embedded networks). Інформація моделюється як мережа, що складається з вузлів та зв'язків між ними. Експертні методи аналізу є доволі трудомісткими. Зважаючи на можливості аналізу з використанням інформаційних технологій, розглянемо більш детально формальні методи аналізу тексту, що можуть бути використанні для створення автоматизованих систем визначення авторства.

Класична задача класифікації передбачає наявність бази даних з певними атрибутами об'єктів, що використовуються як ознаки (features) в алгоритмах машинного навчання. Існує певна специфіка в аналізі текстів, адже база даних зазвичай містить самі тексти, а ознаки (features) отримуються на основі попереднього аналізу перед застосування алгоритму класифікації. Існує ряд проблем, які необхідно розв'язати на цьому етапі, пов'язаних з великою кількістю ознак, частина з яких трапляється рідко з високою часткою шумів та неважливих даних. Тому на етапі попередньої обробки проводять видалення неінформативних слів, а також заміну близьких за значенням слів – одинаковими. Для скорочення простору ознак в текстових документах, як правило, використовують наступні методи: видалення стоп-слів, стеммінг, аналіз n-грам, приведення регістра. Наступним етапом в аналізі текстових документів є виділення ключових слів, яке проводять статистичними методами або лексичними методами. В основі статистичних методів лежить метрика TF-IDF і її модифікації: $TF = \frac{n_i}{n}$, де n_i – число входжень *i-ого* слова в документів, n – загальна кількість слів документу.

$$IDF = log\frac{|D|}{|D_i|},$$

де $|D|$ – кількість всіх документів в датасеті, $|D_i|$ – кількість документів в датасеті, шо містять i-й термін.

TF-IDF метрика визначається, як добуток показників TF та IDF.

До формальних методів визначення автора документу належать задачі класифікації та кластеризації. Ідентифікацію автора можна проводити будь-якими з відомих класифікаційних методів, таких як класифікатор Байєса, метод k-найближчих сусідів, дерев рішень, або методів кластеризації – k-середніх або ж ієрархічними методами.

Слід зазначити, що ідентифікація автора документу може ускладнюватись, коли існує декілька авторів документу, або ж необхідно дати відповідь на питання, чи допомагав хтось автору у створенні документу.