УДК 004.912

ІНФОРМАЦІЙНА СИСТЕМА ВИЗНАЧЕННЯ АВТОРСТВА ТЕКСТУ НА ОСНОВІ СИНТЕЗУ ФОРМАЛЬНИХ МЕТОДІВ

Качановський П. П., студ. гр. МПІн-181
Науковий керівник: **Скітер І. С.,** кф.-м.н., доцент
Національний університет «Чернігівська політехніка»

Досліджуючи методи аналізу текстів на предмет визначення їх авторства можна виділити дві великі групи методів: експертні і формальні.

Експертні методи – це методи, в яких людина виконує роль і аналізатора, експерта, когось з досвідом і навичками, що дозволяють відрізнити текст одного автора від іншого. Експерт може відрізнити стиль одного автора від іншого проаналізувавши такі показники як кількість мовних зворотів, кількість вставних слів, довжина речень, а якщо текст є рукописним навіть і каліграфія. Проте робота експертів може займати дні та може страждати від людського фактору.

Формальні методи – це методи, що використовують елементи теорії розпізнавання образів, математичної статистики та теорії ймовірностей, алгоритми нейронних мереж і кластерного аналізу та багато інших. Такі методи відрізняються від експертних швидкістю та масштабністю, оскільки одночасно можуть аналізувати великий масив текстів.

Слід зазначити, що навіть задачі атрибуції тексту (визначення авторства), можуть бути різними, а саме ідентифікаційними або діагностичні, тобто ті, що не можуть визначити конкретного автора, але можуть визначити його рідну мову, знання іноземних мов, місце народження та інше [1].

Ідентифікаційні завдання дозволяють здійснити перевірку авторства:

- підтвердити авторство певної особи;
- виключити авторство певної особи;
- перевірити той факт, що автором всього тексту був один і той же чоловік;
- перевірити той факт, що той хто написав текст є його справжнім автором;

Постановка таких задач ґрунтується на тому, що автор може бути визначеним.

Діагностичні задачі – це ті задачі де визначити автора точно неможливо, проте можливо визначити його рідну мову, знання іноземних мов, місце народження та інше.

Розроблювальна система спрямована на вирішення ідентифікаційної задачі атрибуції тексту, тому в подальшому під «атрибуцією тексту» буде матися на увазі саме ідентифікація автора серед множини відомих.

Методи атрибуції дозволяють досліджувати текст на п'яти рівнях: пунктуаційному, орфографічному, синтаксичному, лексико-фразеологічному, стилістичному.

Пунктуаційний рівень допомагає виявити особливості вживання автором знаків пунктуації, характерні помилки.

Орфографічний рівень виявляє характерні помилки в написанні слів.

Синтаксичний рівень дозволяє визначити особливості побудови речень, перевагу тих чи інших мовних конструкцій, вживання часів, активної чи пасивної застави, порядок слів, характерні синтаксичні помилки.

Лексико-фразеологічний рівень визначає словниковий запас автора, особливості використання слів і виразів, схильність до вживання рідкісних і іноземних слів, діалектизмів, архаїзмів, неологізмів, професіоналізмів, арготизмів, навички вживання фразеологізмів, прислів'їв, приказок, «крилатих виразів» і т.д..

Стилістичний рівень дозволяє визначити жанр, загальну структуру тексту, для літературних творів – сюжет, характерні зображальні засоби (метафора, іронія, алегорія, гіпербола, порівняння), стилістичні фігури (градація, антитеза, риторичне питання і т. д.), інші характерні мовні прийоми.

Для виконання задачі атрибуції тексту і експертні, і формальні методи намагаються зробити одне і те саме: визначити авторський стиль та порівняти його з вже відомим. Під “авторським стилем” зазвичай розуміють останні три рівни дослідження тексту [3].

Розроблювальна система повинна поєднувати різні формальні методи атрибуції тексту, та повинна бути протестована на текстах різних розмірів, жанрів та стилів. Таке тестування дасть можливість бути впевненим в валідності системи для її подальшого використання або імплементації в вже існуючі.

Одним з основних методів, що будуть використані в розроблювальній системі є так звані інформаційні портрети тексту [2]. Ідея інформаційних портретів полягає у тому, що “авторський стиль” на високому рівні представлений поєднанням слів. Схильність автора використовувати ті чи інші комбінації слів найкраще відображає його стиль і як результат саме комбінації слів слід розглядати для визначення авторства тексту. Проте комбінацій слів може бути велика кількість, один текст може мати від 1000 до 10000 унікальних слів, для визначення частоти використання пар слів необхідно будет побудувати матрицю 10000 на 10000 та проводити її повний аналіз, що є довгим та місцезаратним процесом, натомість інформаційний портрет є матрицею взаємної інформації між комбінаціями літер [4]. Такий підхід скорочує набір даних, що пришвидшує їх накопичення та порівняння. Слід також зазначити, що комбінаціями літер можуть слугувати не тільки сусідні пари літер, а і їх тріади, або пари літер через одну. Найкраще себе може показати система, що проводить побудову і порівняння всіх трьох типів портретів.

Проте може виникнути питання достовірності та доцільності такого підходу. Для тестування цього методу було відібрано 8 авторів XIX-XX століть, для кожного з них було обрано одинадцять творів, десять з яких використовувалися для побудови загального портрету автора, а останній для порівняння з усіма портретами авторів. Як результат з масиву портретів авторів, той, що мав найбільший коефіцієнт кореляції та найменше середнє квадратичне відхилення між матрицею взаємної інформації, обирався як портрет, що належить оригінальному автору тексту. Таке тестування показало, що цей метод дає точність 88**%.**

Список використаних джерел

1. Lomakina L.S., Rodionov V.B., Surkova A.S. Hierarchical Clustering of Text Documents // Automation and Remote Control. 2014. V. 72. № 9. P. 345–351.
2. Ломакина Л.С., Мордвинов А.В., Суркова А.С. Построение и исследование модели текста для его классификации по предметным категориям // Системы управления и информационные технологии. 2011. № 1(43). С. 16–20.
3. Surkova A.S., Domnin A.A., Bulatov I.V., Tsarev A.A. Neural networks and decision trees algorithms – the base of automated text classification and clustering // Am. J. Control Systems and Information Technology. Science Book Publishing House, LLC. 2013. № 2. P. 33–35. 137
4. Smith R.E., Jiang M.K. MILCS: a mutual information learning classifier system // Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '07). 2007. P. 2945–2952.

УДК 330.88(045)

СТВОРЕННЯ СИСТЕМИ КОМУНІКАЦІЙ МІЖ АДМІНІСТРАЦІЄЮ УНІВЕРСИТЕТУ ТА ЗДОБУВАЧАМИ ВИЩОЇ ОСВІТИ

Іскрижицький А. М., студ. гр. ПІ-161,
Іскрижицька О. К., студ. гр. ПІ-161
Науковий керівник: **Трунова О.В**., к.пед.н., доцент
Національний університет «Чернігівська політехніка»

Зміни у системі освіти, зміна поколінь, пріоритетів молодого покоління призводить до зміни взаємодії між адміністрацією університету та здобувачами вищої освіти (ЗВО), традиційна схема спілкування між студентами та адміністрацією стає менш ефективною. Люди звикають до взаємодії через месенджери, мобільні додатки, уникаючи при цьому