ЗАСТОСУВАННЯ МЕТОДІВ СТАТИСТИЧНОГО АНАЛІЗУ ДЛЯ РОЗВ'ЯЗАННЯ ЗАДАЧІ ІДЕНТИФІКАЦІЇ ТЕКСТІВ

Подшиваленко Б.О.
Науковий керівник – к.т.н., доц. Гибкіна Н.В.
Харківський національний університет радіоелектроніки
61166, Харків, просп. Науки, 14, каф. прикладної математики,
тел. (057) 702-14-36, e-mail: borys.podshyvalenko@nure.ua

The aim of this work is research of mathematical methods of identifying the author of the text. Using the method of analysis of hierarchies, the optimal method for solving the problem is chosen according to certain criteria – the method of statistical analysis. The possibility of applying the selected methods to the identification of text of Ukrainian literature is investigated. A software application has been developed, thanks to which it is possible to determine the probability that the text belongs to one or another author from those analyzed.

У наш час, коли інформація в основному зберігається у цифровому вигляді і кожен має доступ до різноманітних інформаційних ресурсів (зокрема, наукових), з'являється можливість вільного і неконтрольованого копіювання цієї інформації. Це стає причиною постійного запозичення авторами окремих виразів, частин документу або й цілих документів, часто без належного цитування, що призводить до втрати достовірності інформації та її цінності. Отже, актуальними стають методи дослідження авторського стилю та встановлення спільного авторства різних текстів.

Для аналізу схожості декількох текстів використовуються спеціалізовані підходи. Аналіз здійснюється на різних рівнях: стилістичному, пунктуаційному, лексико-фразеологічному, орфографічному та синтаксичному. Кожен з цих рівнів дозволяє отримати певну інформацію про структурні особливості досліджуваного тексту та його автора [2].

Велику частку методів аналізу текстів складають математичні методи, зокрема, статистичні методи ідентифікації, на основі яких можна робити висновки про оригінальність розглянутих текстів.

У роботі розглядаються можливості методів статистичного аналізу для ідентифікації текстів за авторами або асоціювання тексту з деяким автором із заданої множини розглядуваних авторів. Аналіз проводиться для творів української художньої літератури декількох відомих авторів. Як узагальнююча характеристика авторського стиля може бути обраний деякий числовий показник, що обчислюється на основі тексту окремого автора та повинен відображати індивідуальний авторський стиль [1, 2]. Порівняння показників, обчислених для двох різних творів, або показника, обчисленого для твору, що досліджується, з еталонним показником певного автора, дає змогу зробити висновок про авторство окремого тексту або про спільне авторство двох досліджуваних текстів.

Задача ідентифікації автора може бути подана у наступному вигляді.

Розглядається множина текстів $T = \{t_1, ..., t_k\}$ і множина авторів $A = \{a_1, ..., a_n\}$. Для деякої підмножини текстів $T' \subseteq T$ автори з множини A відомі. Необхідно визначити, хто з авторів множини A є автором певного тексту з множини текстів $T'' = \{t_{|T'|+1}, ..., t_k\} \subseteq T$ або порівняти два тексти з множини T'' на предмет спільності їх автора.

Для розв'язання поставленої задачі у роботах [1, 2] пропонується визначити співвідношення:

$$Q^{(1)} = \sum_{i=1}^{n} | f_{L_1}(i) - f_{L_A}(i) | \quad (1)$$

або

$$Q^{(2)} = \sum_{i=1}^{n} | f_{L_1}(i) - f_{L_2}(i) | . \quad (2)$$

Тут $f_L(i) = \frac{k_i}{L}$, $i = 1,2,...,n$ – частота виникнення елементу i у тексті довжини L, а k_i – кількість елементів цього типу у наведеному тексті (окремих символів або буквосполучень довжини N кожне, тобто N-грам) [1]. Загальна кількість N-грам мови, якою написані тексти, що аналізуються, дорівнює $n = K^N$, де K – довжина алфавіту.

Значення L – це кількість усіх символів тексту за винятком пунктуаційних знаків, технічних символів, пропусків.

Порівнюючи за допомогою формули (1) характеристику $f_{L_1}(i)$, $i = 1,2,...,n$, тексту довжини L_1 з множини T'' з певним еталонним значенням $f_{L_A}(i)$, $i = 1,2,...,n$, розрахованим для автора A на основі його творів з множини T' загальною довжиною L_A, можна оцінити «близькість» $Q^{(1)}$ авторського стилю досліджуваного тексту та обраного авторського стилю з A. Використовуючи формулу (2) можна порівняти «схожість» $Q^{(2)}$ авторських стилів двох текстів довжини L_1 та L_2 з множини T''. Якщо отримане значення $Q^{(1)}$ (або $Q^{(2)}$ відповідно) не перевищує встановлений поріг, можна зробити висновок про те, що обидва розглядувані твори (або твір і еталонний текст) мають одного й того ж автора.

Обчислювальні експерименти показали придатність описаного методу для аналізу художніх творів української літератури.

Список використаних джерел:

1. Борисов Л. А. Орлов Ю. Н. Осминин К. П. Идентификация автора текста по распределению частот буквосочетаний // Прикладная информатика. 2013. Т. 26. № 2. С. 95-108.

2. Батура Т. В. Формальные методы определения авторства текстов // Информационные технологии. 2012. Т. 10, № 4. С. 81-94.