Журнал обчислювальної та прикладної математики **2020, № 1 (133)** Journal of Numerical & Applied Mathematics

УДК 519.23
MSC 47N30

NONPARAMETRIC METHODS OF AUTHORSHIP ATTRIBUTION IN ENGLISH LITERATURE

D. A. Klyushin, V. Yu. Mykhaylyuk
Faculty of Computer Science and Cybernetics, Taras Shevchenko Kiev National University, Kiev, Ukraine, E-mail: dokmed5@gmail.com

НЕПАРАМЕТРИЧНІ МЕТОДИ АТРИБУЦІЇ АВТОРСТВА В АНГЛІЙСЬКІЙ ЛІТЕРАТУРІ

Д. А. Клюшин, В. Ю. Михайлюк
Факультет комп'ютерних наук та кібернетики, Київський національний університет імені Тараса Шевченка, Київ, Україна, E-mail: dokmed5@gmail.com

Abstract. The paper describes the results of comparison of two nonparametric methods of authorship identification in English literature. It describes testing methods with and without clustering. A method was also proposed to select the n-grams that would best serve as a marker to identify the author. More than 800 texts of 16 authors were used for testing. The method using the density of the distribution is suitable for identifying authors of both large texts (50000+ characters) and small (10000+ characters) ones. A method that uses p-statistics is only suitable for large texts.
Keywords: Text Attribution, Authorship Identification, Petunin Statistics, Clustering, Nonparametric Test.

Анотація. У статті описані результати порівняння двох непараметричних методів ідентифікації невідомого автора на прикладах англійської літератури. У ній описано реалізацію методу кластеризації та застосування методів тестування з кластеризацією та без неї. Запропоновано метод вибору n-грам, які є кращими маркерами для ідентифікації автора. Для тестування було використано понад 800 текстів 16 авторів. В результаті було встановлено, що метод, який використовує щільність розподілу, придатний для ідентифікації авторів як великих текстів (50000 + символів), так і малих (10000+ символів). Метод, який використовує p-статистику, придатний тільки для великих текстів. За допомогою кластеризації текстів на тестовій вибірці для обох методів була досягнута значно кращі результати. Робота продовжує дослідження ефективності методів ідентифікації авторства, виконану раніше на прикладі творів класичної російської літератури. Результати підтверджують, що ефективність методів не залежить від вибраної мови твору.
Ключові слова: атрибуція тексту, ідентифікація авторства, непараметричні методи, статистика Петуніна, кластеризація.

1. Вступ

Використання буквосполучень (n-грам) як стилістичної ознаки тексту для розпізнавання автора вперше запропонував Kjell [1, 2]. В подальшому цей напрямок розвивали Stanatos [3–5], Juola [6], Орлов, Осмінін та інш. [7–9], Дюрдева та ін. [10], Peng et al. [11], Keselj et al. [12], Boughaci [13], Ярошевський та Клюшин [14] та багато інших дослідників. В цих роботах розглянуті різні подходи до автоматичної ідентифікації авторства і проведені численні тести для оцінки точності і ефективності запропонованих методів для атрибуції літературних текстів, написаних різними мовами.

У даній роботі проводиться порівняльний аналіз двох підходів для атрибуції текстів. Перший грунтується на порівнянні середньозважених частотних характеристик буквосполучень (n-грам) [7–9]. Другий метод полягає у перевірці статистичної гіпотези про належність тексту до певного корпусу текстів за допомогою міри неоднорідності розподілів n-грам в навчальних та тестових вибірках [14]. Оскільки в роботі [14] розглядалися твори класичної російської літератури, окремий інтерес викликає питання, наскільки точність методу залежить від особливостей мови, зокрема англійської.

Автори літературних творів можуть змінювати стиль, тому використання одного еталону для автора може бути неефективним. З цієї причини в роботі додатково була реалізована попередня кластеризація текстів автора із подальшим знаходженням еталону для кожного кластеру. Припускається, що такий еталон відповідає окремому стилю автора. Кластеризація виконувалась ієрархічним методом з використанням порогової відстані чи міри однорідності як параметра. Головною метою данної роботи є порівняння методів з точки зору точності і складності реалізації.

2. Ідентифікація автора за середньозваженою частотною характеристикою

Розглянемо метод, описаний в роботі [7]. Уведемо наступні позначення: A — розмір навчальної вибірки текстів, K_α — кількість начальних вибірок автора α, $N_{i,\alpha}$ — кількість букв в i-ому тексті автора α, $f_{i,\alpha}(j)$ — частота j-ї n-грами i-го тексту автора α, де аргумент j змінюється від 1 до 26^n, $N_\alpha = \sum\limits_{i=1}^{K_\alpha} N_{i,a}$ — кількість букв в усіх текстах автора α в навчальній виборці. Для кожного автора введемо сереньозважену частотну характеристику [7]:

$$F_\alpha(j) = \frac{1}{N_\alpha} \sum_{i=1}^{K_\alpha} f_{i,\alpha}(j) N_{i,\alpha}$$

Введемо відстань ρ_{ik} між частотними характеристикакми текстів i та k:

$$\rho_{ik} = \sum_{j=1}^{26^n} \mid f_i(j) - f_k(j) \mid$$

Відповідно до методу [7], для кожного автора α будується щільність функцій розподілу $g_\alpha^+(\rho)$ відхилень $\rho_{i_\alpha,\alpha}$ середньої частотної характеристики тексту, що перевіряється, від текстів автора з навчальної вибірки, а також

щільність розподілу $g_\alpha^-(\rho)$ відповідних відхилень від текстів інших авторів із навчальної вибірки. Позначимо $G_\alpha^\pm(\rho)$ відповідні функції розподілу. Мінімальне значення ρ при якому $G_\alpha^{(\rho)} = 1$, позначимо ρ_α^+, а максимальне значення ρ при якому $G_\alpha^{(\rho)} = 0$, позначимо ρ_α^-.

Робоча гіпотеза [7] полягає в тому що, відстань частотних характеристик усіх текстів автора α від середньої частотної характеристики його текстів не перевищує ρ_α^+, а відстань частотних характеристик усіх текстів інших авторів від середньої частотної характеристики автора α перевищує ρ_α^-. Величина $1 - G_\alpha^+(\rho_\alpha^-)$ — це ймовірність помилки другого роду, а $G_\alpha^-(\rho_\alpha^+)$ – це ймовірність помилки першого роду. Позначимо як $G^+(\rho)$ розподіл відхилень текстів автору від його текстів з навчальної вибірки та як $G^-(\rho)$ розподіл відхилень його текстів від текстів інших авторів

$$G^+(\rho) = \frac{\sum_{\alpha=1}^{A} K_\alpha G_\alpha^+(\rho)}{\sum_{\alpha=1}^{A} K_\alpha},$$
$$G^-(\rho) = \frac{\sum_{\alpha=1}^{A} K_\alpha G_\alpha^-(\rho)}{\sum_{\alpha=1}^{A} K_\alpha}.$$

Пороговою відстанню називається таке значення $\hat{\rho}$, для якої помилка ідентифікації автора тексту є мінімальною.

$$\hat{\rho_\alpha} = \underset{\alpha}{\operatorname{argmin}}(1 - G_\alpha^+(\rho) + G_\alpha^-(\rho)) = \underset{\alpha}{\operatorname{argmax}}(G_\alpha^+(\rho) - G_\alpha^-(\rho))$$
$$\hat{\rho} = \underset{\alpha}{\operatorname{argmin}}(1 - G^+(\rho) + G^-(\rho)) = \underset{\alpha}{\operatorname{argmax}}(G^+(\rho) - G^-(\rho))$$

Величина $\hat{\rho}$ використовується в [7] як верхній рівень кластерізації.

Правило класифікації формулюється за правилом найближчого сусіда відповідно до відхилення від середньої частотної характеристики автора. Якщо це відхилення не перевищує порогової відстані, рішення не приймається.

3. Ідентификація автора за допомогою статистики Петуніна

Описаний нижче метод обгрунтований в роботі [15] і застосований для класифікації текстів класичної російської літератури в [14]. Нехай маємо дві вибірки $x = (x_1, x_2, ..., x_n)$ та $y = (y_1, y_2, ..., y_n)$ з генеральних сукупностей X та Y відповідно. Задача полягає у з'ясуванні, до якої саме генеральної сукупності належить z. Сформулюємо нульову гіпотезу H про однорідність двох вибірок з генеральних сукупностей з функціями розподілу $F_x(u)$ та $F_y(u)$ відповідно, тобто $F_x(u) = F_y(u)$. Тоді, як відомо [16]:

$$p\left(y_k \in (x_{(i)}, x_{(j)}\right) = \frac{j-i}{n+1}, \;\; i < j.$$

Використовуючи вибірку $y = (y_1, y_2, ..., y_n)$, ми можемо знайти частоту h_{ij} випадкової події $y_k \in (x_{(i)}, x_{(j)})$ та розрахувати довірчий інтервал I_{ij} для ймовірності $p(y_k \in (x_{(i)}, x_{(j)})$ із заданим рівнем значущості β. Позначимо L кількість інтервалів для яких виконується $\frac{j-i}{n+1} \in I_{ij}$. Тоді, визначимо міру

однорідності вибірок x та y, як долю інтервалів для яких вірно $\frac{j-i}{n+1} \in I_{ij}$ серед усіх інтервалів:

$$h_{xy} = \frac{2L}{n(n-1)}$$

Оскільки h_{xy} — частота випадкової події $\frac{j-i}{n+1} \in I_{ij}$ з ймовірністю $1-\beta$, ми можемо побудувати довірчий інтервал I_{xy} для події $\frac{j-i}{n+1} \in I_{ij}$ з рівнем значущості β. Якщо $1-\beta \in I$ тоді гіпотеза H приймається, інакше відхиляється. Число h_{xy} є мірою однорідності вибірок x та y. Помінявши x та y місцями та знайшовши частоту h_{yx} і довірчий інтервал I_{yx}, ми можемо побудувати ще один тест для перевірки гіпотези H. Оскільки міра однорідності h_{xy} не симетрична, ми можемо побудувати симетричну міру однорідності:

$$h(x,y) = \frac{1}{2}(h_{xy} + h_{yx}).$$

Нехай маємо навчальну вибірку з K_α творами автора α, де α змінюється від 1 до A. Кожен текст розібьємо на K частин та для кожної частини знайдемо $f^k_{i,\alpha}(j)$. Позначимо $g_{i,\alpha(j)}(j)$ множину частот j-ої n-грами j-ого тексту автора α:

$$g_{i,\alpha}(j) = \{f^1_{i,\alpha}(j), f^2_{i,\alpha}(j), ..., f^K_{i,\alpha}(j)\}.$$

Введемо міру однорідності текстів a та b, як частину n-грам, для яких нульова гіпотеза відхиляється:

$$\|f^{(.)}_a - f^{(.)}_b\| = 1 - \frac{\sum_{j=1}^{a(n)} h(g_a(j), g_b(j)}{26^n}$$

3. Обчислювальні експерименти

Тестування проводилося на сукупності текстів 16 авторів: George Manville Fenn, Sir Walter Scott, R.M. Ballantyne, U.S. Copyright Office, Robert Louis Stevenson, Jules Verne, W.H.G. Kingston, George Sand, Anthony Trollope, Charles Dickens, G. A. Henty, Mor Jokai, Fergus Hume, Alexandre Dumas, E. Phillips Oppenheim, William Le Queux. Навчальна вибірка текстів містіть щонаймеш 50 книг кожного автора довжиною не менше 200000 символів в бібліотеці. Тестування проводилося з використанням 200000 перших символів текстів. Тексти кожного автора було розділено на контрольну та тестову вибірки по 25 текстів в кожній. Оскільки обчислення відстаней для триграм затратне з огляду на обсяг обчислень, були відібрані триграми, які найбільше відрізняють одоного автора від іншого. Для цього була обрана така статистика:

$$v(j) = \frac{\sum_{\alpha=1}^{A} D(f_{\cdot,\alpha}(j))}{D(f_{\cdot,\cdot}(j))},$$

де

$$D(f_{\cdot,\cdot}(j)) = \frac{\sum_{\alpha=1}^{A} \sum_{i=1}^{K_\alpha} (f_{i,\alpha}(j) - M(f_{\cdot,\cdot}(j)))^2}{\sum_{\alpha=1}^{A} K_\alpha},$$

$$M(f_{..}(j)) = \frac{\sum_{\alpha=1}^{A}\sum_{i=1}^{K_\alpha} f_{i,\alpha}(j)}{\sum_{\alpha=1}^{A} K_\alpha}$$

$$D(f_{.\alpha}(j)) = \frac{\sum_{i=1}^{K_\alpha}(f_{i,\alpha}(j) - M(f_{..}(j)))^2}{K_\alpha},$$

$$M(f_{.\alpha}(j)) = \frac{\sum_{i=1}^{K_\alpha} f_{i,\alpha}(j)}{K_\alpha}.$$

Значення $v(j)$ близьке до 0 означає, що при зміні тексту на текст іншого автора частота j-ї n-грами змінюється значно більше ніж при зміні текста на текст того ж автора. Серед усіх 17526 триграм було обрано 1802, для яких $v(j)$ менше 0.6.

Точність ідентифікації автора залежить від кількості частин на які розбивається текст. При збільшенні кількості частин тексту точність ідентифікації за триграмами різко зменшується, що означає що, довжина тексту недостатня для отримання високої точності. При збільшенні кількості частин тексту точність ідентифікації для біграм та монограм зменшується незначно. Для монограм найбільш ефективним виявилося розбиття K=15, для біграм — K=7, для триграм — K=3.

Метод з використанням щільності функції розподілу виявився значно швидшим за метод з використанням p-статистики, за рахунок того, що для p-статистики треба розрахувати велику кількість довірчих інтервалів для кожної n-грами, у порівнянні з суммою різниць частот n-граму. Для монограм, біграм та триграм метод з використанням p-статистики дає кращі результати в точності на $3-4\%$(для текстів довжиною більше 50000). Однак для невеликих за розміром текстів p-статистика дає гірші результати, ніж метод з використанням щільності функції розподілу.

У таблицях 1, 2, 3 міститься інформація про відстані від еталону та відповідні міри однорідності текстів кожного автора для монограм, біграм і триграм відповідно методу [7]. З отриманих таблиц бачимо, що Jules Verne, Charles Dickens та Alexandre Dumas мають досить великі середні відстані до власного еталону, тобто їх стиль різноманітний. Крім того, як відомо, під іменем Alexandre Dumas, часто приховувався колектив авторів із різними стилями.

Перевірка показала, що, починаючи з довжини тексту в 20000 символів, зміна точності ідентификації із збільшення довжини тексту для біграм та триграм э незначною (1-2%). Для монограм таким пороговим значенням є 50000 символів.

Використання кластерізації підвищує точність тестування приблизно на 5% для монограм та приблизно 10% для біграм і триграм у випадку методу з використанням щільності функції розподілу. Найкращу точність 91,75% було отримано для триграм. У випадку методу з використанням статистики Петуніна точність для монограм та триграм зросла приблизно на 5%, а для триграм майже не змінилася. Найкраща точність 85,25% була отримана для біграм.

Автор	Внутрікласова однорідність		Міжкласова однорідність	
	Крос-валідація	Тест	Крос-валідація	Тест
George Manville Fenn	0.033846	0.034025	0.148186	0.160377
Sir Walter Scott	0.028483	0.066368	0.134049	0.147205
R.M. Ballantyne	0.025653	0.025901	0.132092	0.146816
U.S. Copyright Office	0.047425	0.043493	0.230926	0.241641
Robert Louis Stevenson	0.066044	0.050074	0.135040	0.150064
Jules Verne	0.194515	0.232865	0.214297	0.223655
W.H.G. Kingston	0.031068	0.034432	0.140399	0.153805
George Sand	0.034543	0.052155	0.345482	0.354845
Anthony Trollope	0.031290	0.039313	0.144474	0.157774
Charles Dickens	0.137723	0.201273	0.147524	0.156707
G. A. Henty	0.029994	0.031544	0.141307	0.155087
Mor Jokai	0.231570	0.221696	0.266655	0.265208
Fergus Hume	0.034609	0.035530	0.132482	0.146382
Alexandre Dumas	0.101020	0.152490	0.283127	0.291188
E. Phillips Oppenheim	0.027365	0.031932	0.136483	0.150508
William Le Queux	0.031242	0.028961	0.131785	0.145958

Табл. 1. Міри однорідності з еталонами. Монограми.

Автор	Внутрікласова однорідність		Міжкласова однорідність	
	Крос-валідація	Тест	Крос-валідація	Тест
George Manville Fenn	0.099483	0.100713	0.337656	0.364495
Sir Walter Scott	0.085706	0.160734	0.312519	0.337815
R.M. Ballantyne	0.084232	0.083146	0.303563	0.333081
U.S. Copyright Office	0.142319	0.131085	0.518372	0.536536
Robert Louis Stevenson	0.153465	0.126039	0.305945	0.336367
Jules Verne	0.434099	0.507920	0.467984	0.481571
W.H.G. Kingston	0.098637	0.101119	0.319829	0.347821
George Sand	0.095353	0.131873	0.720437	0.736496
Anthony Trollope	0.097833	0.120733	0.335552	0.363841
Charles Dickens	0.289430	0.424478	0.322943	0.339882
G. A. Henty	0.095861	0.097674	0.320820	0.349905
Mor Jokai	0.473379	0.450451	0.564311	0.563759
Fergus Hume	0.103849	0.110262	0.308154	0.336897
Alexandre Dumas	0.228394	0.341310	0.612664	0.625097
E. Phillips Oppenheim	0.085566	0.094804	0.316374	0.344707
William Le Queux	0.092661	0.094849	0.306627	0.334875

Табл. 2. Міри однорідності з еталонами. Біграми.

Автор	Внутрікласова однорідність		Міжкласова однорідність	
	Крос-валідація	Тест	Крос-валідація	Тест
George Manville Fenn	0.180418	0.188610	0.596302	0.623662
Sir Walter Scott	0.171463	0.283575	0.524401	0.549622
R.M. Ballantyne	0.172195	0.169814	0.514450	0.545968
U.S. Copyright Office	0.283822	0.276685	0.815953	0.833536
Robert Louis Stevenson	0.259684	0.228195	0.511306	0.545252
Jules Verne	0.655446	0.770669	0.789989	0.801207
W.H.G. Kingston	0.194357	0.204415	0.539513	0.568905
George Sand	0.162869	0.224804	1.156960	1.169340
Anthony Trollope	0.185929	0.224242	0.572796	0.603712
Charles Dickens	0.466389	0.668956	0.530902	0.546114
G. A. Henty	0.189261	0.192510	0.539592	0.571259
Mor Jokai	0.632626	0.749113	0.671995	0.671159
Fergus Hume	0.192996	0.206572	0.530133	0.561459
Alexandre Dumas	0.368591	0.561702	1.024250	1.029950
E. Phillips Oppenheim	0.170172	0.190420	0.546378	0.576854
William Le Queux	0.194836	0.198912	0.520049	0.550307

Табл. 3. Порівняння відстаней до еталонів. Біграми.

	Навчальна вибірка	Тестова вибірка
Монограми	0.91	0.75
Біграми	0.98	0.89
Триграми	0.99	0.92

Табл. 4. Точність методу з використанням щільності функції розподілу з кластерізацією

	Навчальна вибірка	Тестова вибірка
Монограми	0.97	0.81
Біграми	0.91	0.85
Триграми	0.99	0.81

Табл. 5. Метод з використанням p-статистики. K=15 для монограм, K=7 для біграм, K=3 для триграм

5. Висновки

Порівняння непараметричних методів ідентифікації авторів з використанням середньозваженої частотної характеристики та статистики Петуніна показало, що перший метод гарантує високу точність як для довгих (50000+ символів), так і для коротких текстів (10000+ символів), а другий метод — лише для довгих текстів. Попередня ієрархічна кластерізація текстів дозволяє отримати кращі результати на тестових вибірках для обох методів.

Література

1. Kjell B. Authorship determination using letter pair frequency features with neural network classifiers. *Literary and Linguistic Computing*. 1994. 9(2). P. 119–124.
2. Kjell B., Woods W., Frieder O. Discrimination of authorship using visualization. *Information Processing and Management*. 1994. 30(1). P. 141–150.
3. Stamatatos E. Intrinsic Plagiarism Detection Using Character n-gram Profiles. *In: Benno Stein, Paolo Rosso, Efstathios Stamatatos, Moshe Koppel, and Eneko Agirre, editors, SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09)* Universidad Politecnica de Valencia and CEUR-WS.org, September 2009. P. 38–46.
4. Stamatatos E. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*. 2009. 60(3). P. 538–556.
5. Houvardas J., Stamatatos E. N-Gram Feature Selection for Authorship Identification. *In: Euzenat J., Domingue J. (eds) Artificial Intelligence: Methodology, Systems, and Applications. AIMSA 2006. Lecture Notes in Computer Science*. 2006. vol 4183. Springer, Berlin, Heidelberg, pp. 77–86.
6. Juola P. Authorship attribution. *Found. Trends Inf. Retr.*. 2006. 1(3). P. 233–334.
7. Орлов Ю.Н. Осминин К.П. Определение жанра и автора литературного произведения статистическими методами. *Прикладная информатика*. 2010. Т. 26. № 2. С. 95–108.
8. Орлов Ю.Н. Осминин К.П. Методы статистического анализа литературных текстов. М.: Эдиториал УРСС, 2012.
9. Борисов Л. А. , Орлов Ю. Н., Осминин К. П. Идентификация автора текста по распределению частот буквосочетаний. *Препринты ИПМ им. М. В. Келдыша*. 2013. 027. 26 с.
10. Diurdeva P., Mikhailova E., Shalymov D. Writer identification based on letter frequency distribution. In: *B T. Tyutina, S. Balandin (ed.), 19th Conference of Open Innovations Association*. FRUCT 2016. P. 24–33.
11. Peng J., Choo K., Ashman H. Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles. *Journal of Networked and Computer Applications*. 2016. 70. P. 171–182.
12. Keselj V., Peng F., Cercone N., Thomas C. N-gram-based author profiles for authorship attribution. *Proc. of the Pacific association for computational linguistics*. 2003. 3. P. 255–264.
13. Boughaci D, Benmesbah M., Zebiri A. An improved N-grams based Model for Authorship Attribution. *2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia*. 2019. P. 1–6.
14. Yaroshevskiy A., Klyushin D. Nonparametric Methods of Authorship Attribution in Classic and Modern Literature. In: *2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT), Kyiv, Ukraine*. 2019. PP. 465–469.
15. Klyushin, D.A., Petunin, Yu.I. A Nonparametric Test for the Equivalence of Populations Based on a Measure of Proximity of Samples. *Ukrainian Mathematical Journal*. 2003. 55 (2), P. 181–198.
16. Hill, B.M.. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the ASA*. 1968. 63. P. 677–691.

Надійшла: 20.05.2020 / Прийнята: 18.06.2020

НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ АТРИБУЦИИ АВТОРСТВА В АНГЛИЙСЬКОЙ ЛИТЕРАТУРЕ

Д. А. Клюшин, В. Ю. Михайлюк

Факультет компьютерных наук и кибернетики, Киевский национальный университет имени Тараса Шевченко, Киев, Украина, E-mail: dokmed5@gmail.com

Аннотация. В статье описаны результаты сравнения двух непараметрических методов идентификации неизвестного автора на примерах английской литературы. В ней описана реализацию метода кластеризации и применение методов тестирования с кластеризацией и без нее. Предложен метод выбора n- грамм, которые являются лучшими маркерами для идентификации автора. Для тестирования было использовано более 800 текстов 16 авторов. В результате было установлено, что метод, использующий плотность распределения, подходит для идентификации авторов как больших текстов (50000+ символов), так и малых (10000+ символов). Метод, использующий p-статистику, подходит для использования только в больших текстах. С помощью кластеризации текстов на тестовой выборке для обоих методов была достигнута значительно лучшие результаты. Работа продолжает исследования эффективности методов идентификации авторства, выполненную ранее на примере произведений классической русской литературы. Результаты подтверждают, что эффективность методов не зависит от выбранного языка произведения.

Ключевые слова: атрибуция текста, идентификация авторства, непараметрические методы, статистика Петунина, кластеризация.