РОЗДІЛ V. ПРИКЛАДНА ЛІНГВІСТИКА: НАПРЯМИ Й АСПЕКТИ ДОСЛІДЖЕННЯ

Ірина Волошиновська

ББК Ш11
УДК 81’324’33’37

ЕФЕКТИВНІСТЬ АВТОРСЬКОЇ ТА ТЕМАТИЧНОЇ АТРИБУЦІЇ ТЕКСТІВ НАУКОВО-ТЕХНІЧНОГО СПРЯМУВАННЯ

У даній роботі досліджуються особливості авторської та тематичної атрибуції вузькоспеціалізованих текстів науково-технічного спрямування. Проаналізовано ефективність та особливості застосування методів відносної ентропії та аналізу головних компонент для розділення текстів за авторськими та тематичними ознаками. Атрибуцію текстів проведено на основі аналізу службових слів та n-грам моделей. Виявлено та пояснено залежність функціональних можливостей розглянутих методів атрибуції від розміру n-грам.

Ключові слова: метод аналізу головних компонент, ентропія, службові слова, n-грам модель тексту, авторизація, тематична атрибуція.

Особливості атрибуції спеціалізованих текстів

У випадку робіт з галузі природничих та точних наук автор публікації змушений використовувати усталену термінологію, висловлювати свої ідеї чітко та описувати результати стисло й однозначно. Можливість прояву авторського стилю є досить обмеженою за таких умов. Сучасна стаття науково-технічного спрямування відображає, як правило, результати роботи колективу (до якого інколи активно долучаються рецензенти та редактори журналу) і тому надзвичайно складним є завдання виявлення в ній ознак певного авторського стилю. Значний вклад термінологічної лексики в наукових роботах обумовлює суттєву перевагу тематичних лексичних ознак над ознаками авторського стилю.

Успішне вирішення питань тематичної або авторської атрибуції текстів залежить від наступних двох кроків: 1) коректного вибору параметрів, що відповідають за тематику документів або авторський стиль; 2) пошуку відповідних методів та алгоритмів розділення робіт за тематичними або авторськими ознаками. Параметрами оцінки тематики або авторського стилю тексту можуть виступати фонологічні, морфологічні, лексичні, синтаксичні та структурні властивості тексту. Кількість параметрів для одночасного аналізу текстів є необмеженою. Процеси ідентифікації та діагностики авторського стилю базуються, в основному, на аналізі особливостей використання автором синтаксичних конструкцій та граматичних форм, оцінці об’єму лексичного словника автора, середньої кількості слів у реченні, кількості речень в абзаці тощо [Марусенко 1990; Милов 2009; Перебийніс 1967; Севбо 1981]. Методи тематичної та авторської атрибуції текстів можна поділити на контрольовані (*supervised*) та неконтрольовані (*unsupervised*) [Juola 2006: 272-286]. Контрольовані методи вимагають апріорних даних про досліджуваний текст, а систему, яка працює за даним методом, необхідно навчати для успішного проведення атрибуції. Прикладами контрольованих методів є лінійний дискримінантний аналіз та машина опорних векторів. Неконтрольовані методи не вимагають апріорних даних про текст, а система самостійно виконує поставлену задачу без зовнішнього втручання. До неконтрольованих методів відносять метод аналізу головних компонент, кластерний аналіз. Можливість ефективного застосування методу аналізу головних компонент для авторизації тексту показано в роботах [Burrows 1992; Burrows 2003]. Автори [Baayen 2002; Juola 2003] продемонстрували успішну класифікацію робіт за тематикою, автором, його віковою категорією та освітою, використовуючи метод лінійного дискримінантного аналізу з врахуванням ентропії частотних характеристик слів в межах досліджуваного набору текстів.

Слід зауважити, що і до тепер однією з основних проблем атрибуції залишається відкрите питання щодо спільного консенсусу стосовно методів та параметрів, які використовуються для успішного подолання завдань стилеметрії [Holmes 1998; Jockers 2010], що і зумовлює **актуальність** досліджень, представлених у даній роботі. Методика, що є успішною для одного із завдань атрибуції, може виявитись непридатною для інших завдань, кожен конкретний випадок вимагає свого індивідуального підходу і адаптації алгоритму аналізу [Rudman 1997]. **Метою** даної роботи є оцінка ефективності методів відносної ентропії та аналізу головних компонент для розділення текстів за авторськими та тематичними ознаками.

Методи розрахунків

В даній роботі досліджуються функціональні можливості методу аналізу головних компонент [Jolliffe 2002] та методу відносної ентропії [Kullback 1951] для тематичної та авторської атрибуції наукових праць. Метод аналізу головних компонент полягає у виборі системи координат для оптимального представлення багатомірних даних. Осями такої координатної системи вибираються власні вектори $\mathbf{p}_i$ коваріаційної матриці:

$$\mathrm{cov}(\mathbf{A}) \cdot \mathbf{p}_i = \lambda_i \cdot \mathbf{p}_i \ , \qquad (1)$$

де λ_i – власні значення коваріаційної матриці, яка формується наступним чином:

$$\mathrm{cov}(\mathbf{A}) = (\mathbf{A}^{\mathrm{T}} \cdot \mathbf{A}) / (m-1) \ , \qquad (2)$$

m – кількість стрічок транспонованої матриці $\mathbf{A}^{\mathrm{T}}$ (кількість досліджуваних текстів), а елемент a_{ij} матриці $\mathbf{A}$ відображає ймовірність появи *i*-ого слова в *j*-ому тексті.

Подібність, або ж кореляція особливостей статистичного розподілу лексико-синтаксичних одиниць в текстах може розглядатись як доказ подібності цих текстів за відповідними ознаками. Дисперсія для ймовірності $p(x_j)$ появи одиниці x_j в тексті S визначається ентропією:

$$H(S) = \sum_{j=1}^{n} p(x_j) \times \log_2 \ p(x_j) \qquad (3)$$

Відносна ентропія обчислюється для оцінки подібності текстів S_i до наперед визначеного (опорного) тексту S_1. Така міра подібності текстів (відносна ентропія) відома також як дивергенція Кульбака-Лайблера (*Kullback-Leibler divergence, KLD*):

$$KLD(S_1; S_i) = \sum_{j=1}^{n} p_1(x_j) \times \log_2 \frac{p_1(x_j)}{p_i(x_j)} \ , \qquad (4)$$

де $p_i(x_j)$ - ймовірність появи одиниці x_j в S_i тексті.

Аналіз ефективності згаданих вище методів проведено для *n*-грам моделей [Cavnar 1994] досліджуваних текстів (n=1..10) а також для службових слів. Найбільш вдалою для англомовних текстів вважається 3-грам модель [Lafferty 1992], а особливості використання автором службових слів розглядаються як характерні ознаки авторського стилю в художній літературі [Zhao 2005].

Для проведених в даній роботі досліджень вибрано 40 праць з галузі фізики твердого тіла, опублікованих різними науковими групами англійською мовою в реферованих журналах. Праці відібрані після попереднього вивчення складів та сфер діяльності наукових груп. Тематика проаналізованих робіт дотримана таким чином в межах вузького спрямування у галузі люмінесцентної спектроскопії з метою перевірки роздільної здатності застосованих методів для авторської та тематичної атрибуції. Праці розділені на 4 групи (по 10 в кожній) відповідно до наукових груп та авторів (співавторів): 1) проф. д-р. П. Доренбос (Pieter Dorenbos); 2) проф. д-р. А. Майєрінк (Andries Meijerink); 3) д-р. Григорій Стриганюк (Gregory Stryganyuk); 4) проф. д-р. Г. Ціммерер (Georg Zimmerer). Загальний словник проаналізованих робіт налічує 11385 словоформ.

Завданням даної роботи є з'ясування особливостей та оптимальних умов розділення робіт за тематичними та авторськими ознаками з використанням методів відносної ентропії та аналізу головних компонент.

Результати дослідження та їх обговорення

Дивергенцію Кульбака-Лайблера (4) пораховано для 40 досліджуваних текстів з метою виявлення їх спорідненості в межах наперед окреслених 4-ох груп. На рисунку 1 відображені результати оцінки спорідненості (розбіжності) текстів, отримані методом обчислення відносної дивергенції (4) за службовими словами. До розгляду було взято 360 службових слів [Zhao 2005]. Тексти на рисунку 1 (D*i*, M*i*, S*i*, Z*i*; *i*=1..10) позначені відповідно до першої літери прізвища автора та проранговані по осі абсцис за мірою розбіжності у зростаючому порядку. Тексти пронумеровані в межах кожної групи відповідно до їх розміру в порядку спадання. Відносна ентропія обчислювалась для 4-ох опорних текстів (D1, M1, S1, Z1), які було вибрано з огляду на їх максимальний об'єм в межах відповідної групи.

Роботи групи D виявились найбільш спорідненими за результатами обчислення відносної ентропії для службових слів. З 10-ти робіт групи D лише 7 (70%) отримали ранг <11 відповідно до міри їх розбіжності (рис. 1а). Ідеальний результат атрибуції (100%) для досліджуваних текстів очікується як мінімальна дисперсія відносної ентропії (розбіжності) в межах однієї наперед сформованої групи, тобто коли усі 10 текстів рангуються <11 відповідно до міри їх розбіжності. Лише 6 текстів групи M (60%) рангуються в межах 1-10 (рис. 1б), а про спорідненість текстів груп S та Z важко й здогадуватись (рис. 1в, г), виходячи з результатів обчислення відносної ентропії за службовими словами. Звичайно, результати обчислень могли б бути дещо іншими у випадку вибору інших опорних текстів, однак на практиці така можливість не завжди є. Вибором опорних текстів максимальних об'ємів передбачалось охоплення максимальної кількості характерних ознак, за якими можливо було б об'єднати тексти відповідної групи.

Аналіз рангово-ймовірнісного розподілу слів у текстах художньої та наукової літератури [Волошиновська 2008] вказує на їх суттєві відмінності в області високих рангів, де основний вклад припадає саме на службові слова. Відносно менший вклад службових слів, виявлений для наукової літератури, може бути причиною складнощів атрибуції наукової літератури за службовими словами.

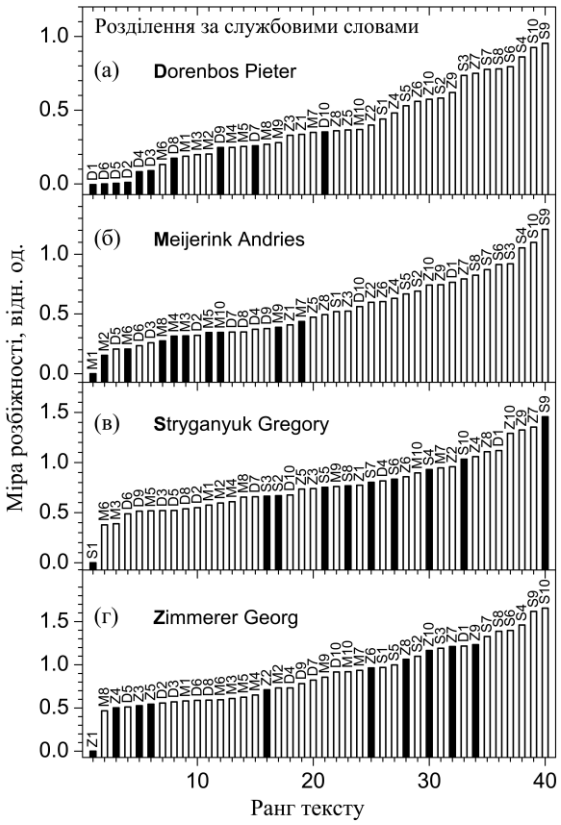

Рис. 1. Розподіл текстів за мірою розбіжності, обчисленою методом дивергенції Кульбака-Лайблера для випадку аналізу службових слів.

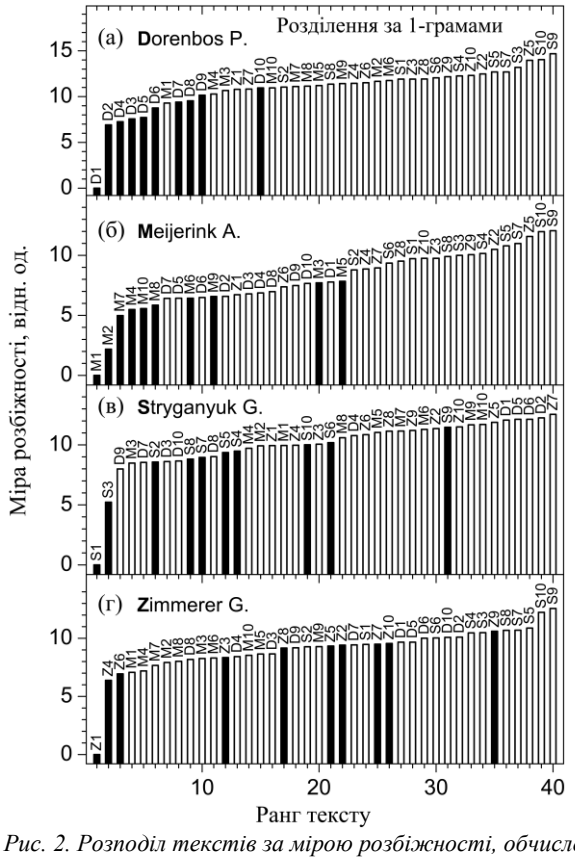

Рис. 2. Розподіл текстів за мірою розбіжності, обчисленою методом дивергенції Кульбака-Лайблера для випадку аналізу 1-грам моделі.

Розрахунки дивергенції Кульбака-Лайблера (КЛД) були проведені також для *n*-грам моделей досліджуваних текстів. На рисунку 2 представлені результати оцінки спорідненості текстів за КЛД значеннями для випадку 1-грам моделі. Атрибуції (групування) 90% текстів досягнуто для групи D (рис. 2а), що є на 20% краще від випадку аналізу службових слів. На 10% покращилась атрибуція для текстів групи M і сягнула 70% (рис. 2б). Однак, групування текстів груп S та Z не спостерігається (рис. 2в, г) і у випадку КЛД обчислень для 1-грам моделі.

Дослідження особливостей авторської атрибуції наукових праць методом аналізу головних компонент *n*-грам моделі текстів [Волошиновська 2009] показали, що атрибуція лише за тематичними ознаками можлива для 1-грам моделі, в той час як авторські ознаки атрибуції стають переважаючими при зростанні розміру *n*-грам до 4. Отже, групування текстів груп D та M для 1-грам моделі (рис. 2) може відбуватись виключно за тематичними ознаками, які слабше проявлятись в групах S та Z через мале тематичне перекриття відповідних праць, обумовлене, наприклад, широким набором об’єктів досліджень у цих групах.

Найбільш чітке групування досліджуваних текстів за значеннями КЛД (4) досягнуто для їх 4-грам моделі (рис. 3). Атрибуції на рівні 70-80% досягнуто для 4-грам моделі текстів груп D та M (рис. 3а, б), групування яких, як виявилось, не сильно чутливе до зміни розміру *n*-грам в межах *n* від 1 до 5. В групі S спостерігається повне групування (рис. 3в; атрибуція 100%). Атрибуції на рівні 60% досягнуто в групі Z (рис. 3г), тексти якої проявляють дуже малу схильності до групування. Слід зауважити, що атрибуція робіт групи S за тематичними ознаками (1-грам модель) виявилась практично неможливою, однак спорідненість робіт групи S була чітко виявлена за авторськими ознаками (4-грам модель).

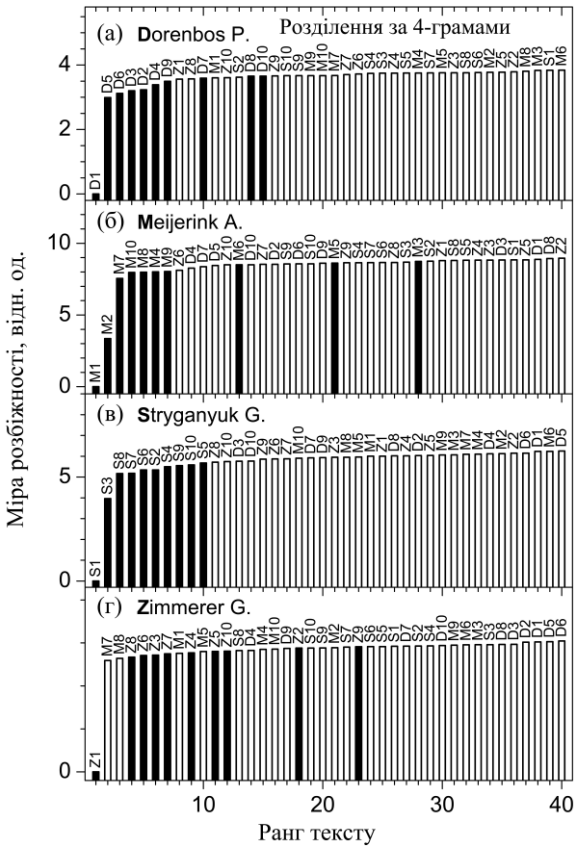

Рис. 3. Розподіл текстів за мірою розбіжності, обчисленою методом дивергенції Кульбака-Лайблера для випадку 4-грам моделі.

Залежність об'єму загального словника досліджуваних текстів від розміру *n*-грам приведено на рисунку 4. Об'єм словника сягає свого максимального значення (107349) саме у випадку 4-грам моделі текстів, для якої й було досягнуто найкращого групування досліджуваних текстів (рис. 3). Те ж саме значення оптимального розміру *n*-грам було виявлене при авторській атрибуції вузькоспеціалізованих наукових праць методом аналізу головних компонент [Волошиновська 2009], де зроблено висновок щодо оптимального розміру *n*-грам, при якому досягається максимальний об'єм словника та забезпечується найбільший прояв ознак авторського стилю.

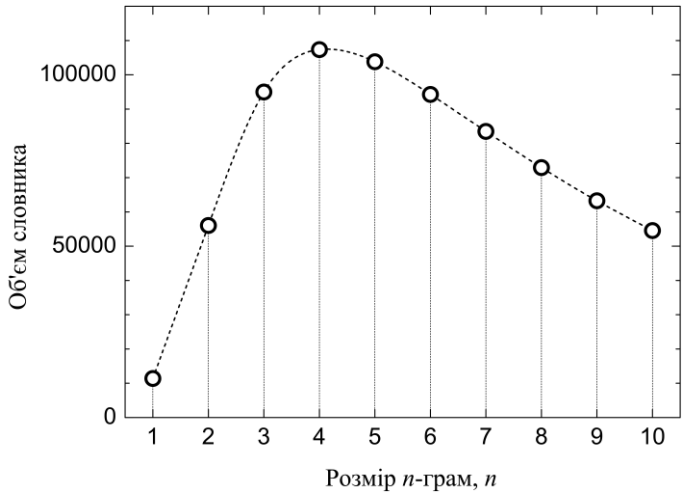

Рис. 4. Залежність об'єму словника від розміру n-грам.

На рисунку 5 представлено результат атрибуції досліджуваних текстів методом аналізу головних компонент їх 4-грам моделі. Розподіл статей представлено в системі координат головних компонент (*Principal Component*) PC-3..5, які описують 9.45% дисперсії вхідних даних, чого виявилось достатньо для чіткого

розділення досліджуваних робіт на відповідні їм групи D, M, S та Z. Групи S, D, а також пара груп M-Z розділяються вздовж осі PC-3. Групи M та Z розділяються між собою за характеристиками, що формують головну компоненту PC-5. Ці групи (M та Z) розведені одна від одної вздовж осі PC-5. Роботи групи D розподілені, в основному, вздовж осі PC-4. Як і у випадку аналізу дивергенції Кульбака-Лайблера, метод аналізу головних компонент дозволив чітко окреслити роботи групи S, та виявити перекриття груп D, M та Z, роботи яких проявляють спорідненість також за близькими значеннями відносної ентропії.

Детальний аналіз об'єктів досліджень робіт дозволив пояснити виявлене перекриття груп текстів наявністю спільної термінології (об'єктів, характеристик) для різних груп. Навіть чітко відокремлена група S включає в себе роботи S5 та S9, які наближаються до області перекриття груп D, M та Z (рис. 5). Причиною такого прояву спорідненості робіт S5 та S9 з іншими групами є наявність спільних об'єктів (кристалічних матриць $LaCl_3$ та $LiLuF_4$). Отже, навіть у випадку 4-грам моделі, вклад тематичних ознак є значним і характер атрибуції наукового-технічного тексту не може бути чітко визначеним. Обидва розглянуті методи забезпечують гібридизовану тематично-авторську атрибуцію науково-течнічних текстів навіть у випадку їх 4-грам моделі. Роздільна здатність методу дивергенції Кульбака-Лайблера є, щоправда, значно обмеженою через його одномірність. Метод аналізу головних компонент забезпечує можливість моніторингу лексики, що визначає характеристики аналізованих текстів. Кількість таких характеристик визначається кількістю головних компонент моделі.

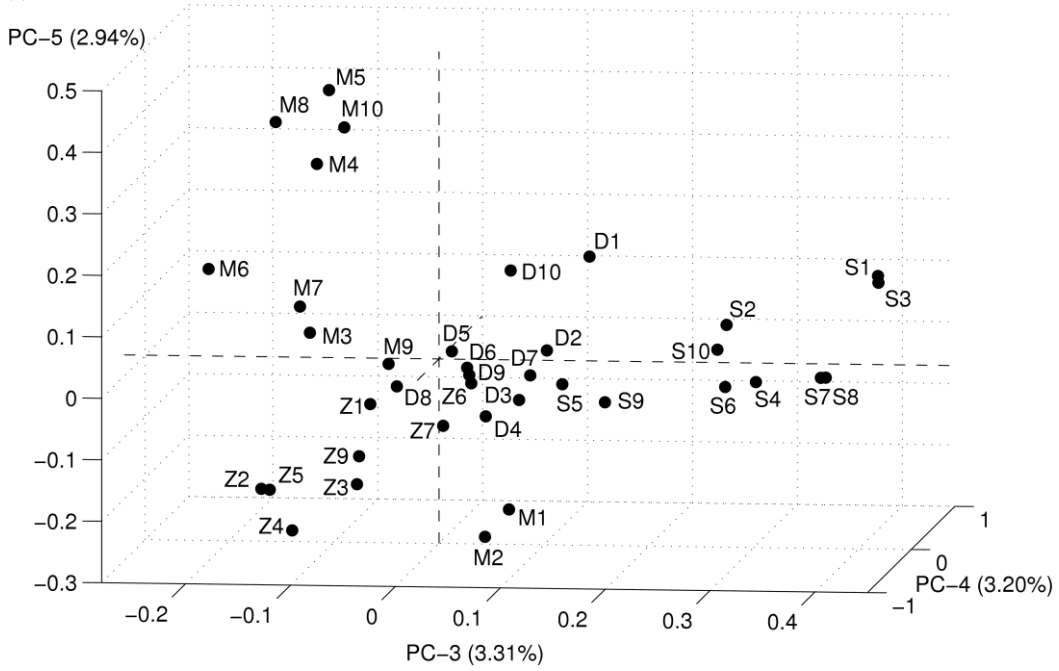

Рис. 5. Розподіл текстів у просторі головних компонент (PC-3..5) для випадку 4-грам моделі.

Висновки

Застосування розглянутих методів аналізу головних компонент та дивергенції Кульбака-Лайблера (КЛД) до *n*-грам моделі наукових текстів дозволяє проводити їх тематичну та авторську атрибуцію з можливістю перерозподілу ваги тематичних та авторських ознак шляхом зміни розміру *n*-грам. Ознаки авторського стилю більше проявляються при збільшенні розміру *n*-грам, однак значний вклад чітко виражених тематичних ознак наукової праці обумовлює гібридизований тематично-авторський характер атрибуції. Оптимальна ефективність такої атрибуції досягається при розмірі *n*-грам, для якого об'єм словника сягає свого максимального значення. Метод аналізу головних компонент, на відміну від КЛД методу, не вимагає апріорних даних, є багатомірним, дозволяючи розділення кількох характеристик одночасно і забезпечує детальну інформацію про вклад лексичних одиниць у відповідні характеристики. Метод КЛД вимагає задання опорного тексту і забезпечує одномірне представлення результатів атрибуції. В перспективі планується перевірка ефективності розглянутого в даній роботі поєднання *n*-грам моделі з методами КЛД та аналізу головних компонент для текстів українською мовою шляхом їх апробації на повідомленнях електронної пошти, творах (переказах) студентів, статтях електронних видань.

Література

Волошиновська 2008: Волошиновська, І. А. Модифікація функції розподілу Лавалетті як адаптація рангово-частотного закону Зіпфа для текстового корпусу природної мови [Текст] / І. А. Волошиновська // Лінгвістичні студії: Збірник наукових праць. – 2008. – №16. – С. 334-339.

Волошиновська 2009: Волошиновська, І. А. Особливості авторизації вузькоспеціалізованих наукових праць [Текст] / І. А. Волошиновська // Нова Філологія: Збірник наукових праць. – 2009. – №35. – С. 36-43.

Марусенко 1990: Марусенко, М. А. Атрибуция анонимных и псевдонимных литературных призведений методами распознавания образов [Текст] / М. А. Марусенко, – Л.: Издательство Ленинградского университета, 1990. – 168 с.

Милов 2009: От Нестора до Фонвизина. Новые методы определения авторства [Текст] / Л. В. Милов (ред.), Л. И. Бородкин, Т. И. Иванова и др. – Магадан, 2009. – 448 с. – ISBN 978-5-244-00291-1.

Перебийніс 1967: Перебийніс, В. С. Статистичні параметри стилів [Текст] / В. С. Перебийніс (ред.). – К.: Наукова Думка, 1967. – 260 с.

Севбо 1981: Севбо, И. П. Графическое представление синтаксических структур и стилистическая діагностика [Текст] / И. П. Севбо, – К.: Наукова Думка, 1981. – 372 с.

Baayen 2002: Baayen, H., van Haltern, H., Neijt, A., Tweedie, F. An experiment in authorship attribution, in Proc. of *JADT* [Text] / H. Baayen, H. van Haltern, A. Neijt, F. Tweedie. – 2002, St.Malo, pp. 29-37.

Burrows 1992: Burrows, J. Not unless you ask nicely: The interpretative nexus between analysis and information, *Literary and Linguistic Computing*, Vol. 7 [Text] / J. Burrows, 1992, pp. 91-109.

Burrows 2003: Burrows, J. Questions of Authorships: Attribution and Beyond, *Computers and the Humanities*, Vol. 37, No. 1 [Text] / J. Burrows, 2003, pp. 5-32.

Cavnar 1994: Cavnar, W. B., Trenkle, J. M. N-Gram-Based Text Categorization, in Proc. of 3rd Annual Symposium on Document Analysis and Information Retrieval, Las Vegas [Text] / W. B. Cavnar, J. M. Trenkle, 1994, pp. 161-175.

Holmes 1998: Holms, D. The Evolution of Stylometry in Humanities Scholarship, *Literary and Linguistic Computing*, Vol. 13, No. 3 [Text] / D. Holms, 1998, pp. 111-117.

Jockers 2010: Jockers, M. L., Witten, D. M. A comparative study of machine learning methods for authorship attribution, *Literary and Linguistic Computing*, Vol. 25, No 2 [Text] / M. L. Jockers, D. M. Witten, 2010, pp. 215-223.

Jolliffe 2002: Jolliffe, I. T. Principal Component Analysis (Springer Series in Statistics) 2nd ed., NY: Springer [Text] / I. T. Jolliffe, 2002, 487 p.

Juola 2003: Juola, P. The time Course of language change, *Computers and the Humanities*, Vol. 37, No. 1 [Text] / P. Juola, 2003, pp. 77-96.

Juola 2006: Juola, P. Authorship attribution, *Foundations and Trends in Information Retrieval*, Vol 1, No 3 [Text] / P. Juola, 2006, pp. 233-334.

Kullback 1951: Kullback, S., Leibler, R. A. On information and sufficiency, Annals of Math. Stat., Vol. 22 [Text] / S. Kullback, R. A. Leibler, 1951, pp. 79-86.

Lafferty 1992: Lafferty, J. D., Sleator, D., Temperley, D. Grammatical trigrams: A probabilistic model of link grammar, in Proc. of AAAI Fall Symp. Probabilistic Approaches to Natural Language, Cambridge, MA, Oct. 1992 [Text] / J. D. Lafferty, D. Sleator, D. Temperley, pp. 74-81.

Rudman 1997: Rudman, J. The State of Authorship Attribution Studies: Some Problems and Solutions, *Computers and the Humanities*, Vol. 31 [Text] / J. Rudman, 1998, pp. 351-365.

Zhao 2005: Zhao, Y., Zobel J. Effective authorship attribution using function words // *Proceedings of the 2nd AIRS Asian Information Retrieval Symposium*, Jeju Island, South Korea, Springer [Text] / Y. Zhao, J. Zobel 2005, pp. 174-190.

В работе исследуются особенности авторской и тематической атрибуции научно-технических работ. Проанализированы эффективность и особенности применения методов относительной энтропии и анализа главных компонентов для разрешения текстов по авторским и тематическим признакам. Проведено атрибуцию текстов на основании анализа служебных слов и моделей n-грамм. Обнаружена и объяснена зависимость функциональных возможностей рассмотренных методов атрибуции от размера n-грамм.

Ключевые слова: метод анализа главных компонентов, энтропия, служебные слова, n-грамм модель текста, авторизация, тематическая атрибуция.

Present work reports on the peculiarities of authorization and thematic attribution revealed for scientific texts. The efficiency of attribution has been estimated for the principal components' analysis and relative entropy techniques. The attribution of the considered texts has been implemented via analysis of functional words and n-gram models. The dependence of functional capabilities on n-gram size has been revealed for the considered attribution techniques.

Keywords: principal components' analysis, entropy, functional words, n-gram model of text, authorization, thematic attribution.

Надійшла до редакції 8 вересня 2010 року.